JOHANNES KEPLER
UNIVERSITÄT LINZ | JKU

TNF

Technisch-Naturwissenschaftliche
Fakultät

All-at-once Multigrid Methods for Optimality Systems Arising from Optimal Control Problems

DISSERTATION

zur Erlangung des akademischen Grades

Doktor

im Doktoratsstudium der

Technischen Wissenschaften

Eingereicht von:
Dipl.-Ing. Stefan Takacs

Angefertigt am:
Doktoratskolleg Computational Mathematics

Beurteilung:
Prof. Dr. Walter Zulehner, Johannes Kepler Universität Linz (Betreuung)
Prof. Dr. Andrew Wathen, University of Oxford

Linz, August 2012

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Die vorliegende Dissertation ist mit dem elektronisch übermittelten Textdokument identisch.

Linz, im August 2012

Stefan Takacs

Acknowledgments

This thesis covers results that where achieved within my research at the Doctoral Program "Computational Mathematics" at the Johannes Kepler University Linz.

First of all I own my deep gratitude to my supervisor Walter Zulehner for advising my thesis, and all the countless discussions, inspirations and helpful hints given throughout the last years. He has raised many of the interesting questions discussed in this thesis. Nonetheless, he gave me the opportunity to work also on problems I have posed myself.

At the same time I would like to thank Andy Wathen for showing much interest in my work and for co-refereeing this thesis. Moreover, I thank him, as well as Volker Schulz and Roland Herzog, for giving me the opportunity to visit their research groups at the University of Oxford, the Universität Trier and the TU Chemnitz.

Special thanks go to Veronika Pillwein for her contributions to the joint work on local Fourier analysis, especially her help in learning more about symbolic computation and for various discussions about my work.

My personal thanks go to my family for their constant understanding and support, in particular to my parents for being role-models and for bringing me in contact in Mathematics and for arousing my interest in that topic.

I gratefully acknowledge the scientific environment at both the Institute of Computational Mathematics and the Doctoral Program "Computational Mathematics", and I want to thank my colleagues there for their hospitality. The discussions with colleagues, especially at the regular meetings of the Doctoral Program in Strobl and Linz, led to fruitful discussions and helpful hints.

Last but not least I thank the Austrian Science Funds (FWF) for funding my work under grant W1214-N15, project DK12.

Zusammenfassung

In dieser Arbeit konstruieren und analysieren wir Mehrgittermethoden zur Lösung gewisser Klassen von Optimalsteuerungsproblemen. Ursprünglich wurden Mehrgittermethoden zur Lösung elliptischer Probleme konstruiert. Die Optimalsteuerungsprobleme jedoch werden durch ein lineares System notwendiger Bedingungen beschrieben (Optimalitätssystem), das nicht elliptisch ist. Wir machen uns jedoch zunutze, dass das Optimalitätssystem eine Block-Matrix ist, das eine Sattelpunktstruktur aufweist: Einerseits haben wir die beiden Blöcke von Variablen, die bereits Teil des Optimalsteuerungsproblems sind: die Variablen, die den Zustand beschreiben, und die Variablen, die die Kontrolle beschreiben. Ferner bilden die Lagrange-Multiplikatoren einen dritten Block von Variablen, der beim Übergang zum Optimalitätssystem eingeführt wird.

Es gibt nun mehrere Möglichkeiten, in diesem Kontext Mehrgitterverfahren zu verwenden. Eine Möglichkeit besteht darin, das Mehrgitterverfahren als Teil eines Vorkonditionierers jeweils auf einzelne Blöcke des Gesamtsystems anzuwenden. Dazu müsste die Mehrgittermethode in jedem Schritt des jeweils gewählten äußeren Iterationsverfahrens angewandt werden. Eine andere Möglichkeit, der wir in dieser Arbeit folgen, ist es, die Methode direkt auf das Gesamtsystem anzuwenden. Ein solcher Zugang wird auch als *all-at-once approach* bezeichnet.

Für einen solchen Ansatz ist der wesentlichse Punkt die Wahl eines passenden Glätters. Wir werden Glätter konstruieren, deren Konvergenzraten vom Grad der Verfeinerung der Diskretisierung unabhängig sind. Da in diesem Fall der Gesamtaufwand der Methode linear von der Anzahl der Unbekannten abhängt, sprechen wir auch von einer optimalen Konvergenzrate.

Für eine Teilklasse der betrachteten Probleme gehen wir einen Schritt weiter und konstruieren auch Lösungsmethoden, deren Konveregnzraten robust in einem Kosten- oder Regularisierungsparameter sind, der Teil der Problemstellung ist. Methoden, die auf eine solche Robustheit nicht Rücksicht nehmen, zeigen für kleine Werte dieses Parameters typischerweise sehr langsame Konveregenz.

Die Konvergenztheorie wird auf einen allgemeinen Konvergenzsatz aufgebaut, der auf eine große Klasse von Methoden anwendbar ist. Der Konvergenzbeweis selbst folgt klassischen Ideen und ist auf der von Hackbusch eingeführten Aufspaltung in Glättungs- und Approximationseigenschaft aufgebaut. Danach wenden wir noch eine zweite Art der Konvergenzanalyse an: lokale Fourieranalyse. Dieser Ansatz erlaubt uns, scharfe Abschätzungen der Konveregenzrate zu bestimmen.

Abstract

In this thesis we construct and analyze multigrid methods for solving the optimality system characterizing the solution of an optimal control problem. Originally multigrid methods were constructed for elliptic problems. However, the (discretized) optimality system is not elliptic. We make use of the fact that the matrix is a block-matrix with saddle point structure: on the one hand we have two blocks of variables representing state and control. These two blocks are already part of the optimal control problem itself. On the other hand, the conversion to the optimality system requires the introduction of Lagrange multipliers, which form the third block of variables.

There are several possibilities to use multigrid methods for constructing solvers for such saddle point problems. One approach to solve such problems is to apply multigrid methods in every step of an overall block-structured iterative method to equations in just one of these blocks of variables. Another approach, which we will follow here, is to apply the multigrid idea directly to the (reduced or not reduced) optimality system, which is called an all-at-once approach.

The choice of an appropriate smoother is the key issue in constructing such a multigrid method. The other part of the method – the coarse-grid correction – can be set up in a canonical way because we will use a framework of conforming geometric multigrid method. In this framework the smoother will be constructed such that the convergence rates are independent of the grid level. This leads to an overall computational complexity that is linear in the number of unknowns which is called an optimal convergence.

For a sub-class of the class of problems we will introduce in a first point, we will go one step further and construct methods where the convergence rates are also robust in a certain regularization or cost parameter which is part of the problem. Moreover, we will show this fact. Methods that do not take this into account typically show quite poor convergence rates if this parameter attains small values.

For the analysis, we will introduce a general framework that is based on Hackbusch's splitting of the analysis into smoothing and approximation property. This allows to give general convergence theorems for the methods under investigation. A second point we consider is local Fourier analysis which allows to compute sharp bounds of the convergence rates.

Contents

Chapter 1

Introduction

The goal of this thesis is the construction and the analysis of fast numerical methods for computing approximate solutions for optimization problems with partial differential equations as constraints (PDE-constrained optimization problems). Problems which belong to this class are optimal control problems, (cf. Lions [42] and Tröltzsch [65]) optimal design problems, shape and topology optimization problems (cf. Bendsøe and Sigmund [7], Pironneau [50]) and many others. In this thesis we focus on optimal control problems. Often, optimal control problems have a quadratic goal functional and linear constraints. Such problems are called quadratic optimization problems. If other problems are considered, typically linearization techniques, like (semi-smooth) Newton methods, lead to such subproblems. Also problems with additional algebraic inequality constraints can be approximated by a sequence of quadratic optimization problems, see, e.g., Ito and Kunisch [40], Gfrerer [32], Herzog and Sachs [37] and others. Therefore the construction of fast solvers for quadratic PDE-constrained optimization problems is of particular interest.

In principle, black-box methods for solving such optimization problems are possible, i.e., one can use appropriate solvers for the PDEs forming the constraints and construct an appropriate outer iteration for solving the optimization problems. Typically, already the solution of the PDE itself is relatively costly, especially if a fine resolution of the problem is required. Therefore, alternative approaches which directly lead to the solution of the optimization problem are of particular interest.

We use the fact that the solution of the problems of our interest can be characterized by their optimality systems, which are also called Karush-Kuhn-Tucker systems (KKT-systems). In this thesis we only consider quadratic problems. For this case, the first order optimality system is a necessary and sufficient condition for a solution.

After discretization, the optimality system is a large-scale linear system which has saddle point structure. We will study multigrid methods for solving these linear systems. Originally, multigrid methods have been designed and analyzed for elliptic problems. They also work well for saddle point problems (like the KKT-systems for PDE-constrained optimization and particularly optimal control problems) and have gained growing interest in this area, see, e.g., Borzi and Schulz [13] and the references cited there. Neither the construction of such multigrid solvers, nor their analysis is standard.

The unknowns of the discretized KKT-system of a PDE-constrained optimization problem can be partitioned into primal and dual variables. For optimal control problems the primal unknowns are the state variable and the control variable. The dual variables are Lagrange multipliers that are introduced to incorporate the constraints into the optimality system. One approach to solve such problems is to apply multigrid methods in every step of an overall block-structured iterative method to equations in just one of these blocks of variables. Such methods have been proposed, e.g., in Hackbusch [34], Battermann and Heinkenschloss [5], Battermann and Sachs [6], Biros and Ghattas [11, 10], Hazra and Schulz [36], Schöberl and Zulehner [54], Zulehner [71] and Rees, Dollar and Wathen [51].

Another approach, which we will follow here, is to apply the multigrid idea directly to the (reduced or not reduced) KKT-system. This approach is called an all-at-once approach. Such methods have been proposed, e.g., in Taasan [60], Arian and Taasan [2], Trottenberg [66], Borzi, Kunisch and Kwak [12], Schulz and Wittum [55], Borzi and Schulz [13], Lass [41], Simon and Zulehner [58] and Schöberl, Simon and Zulehner [53].

The choice of an appropriate smoother is a key issue in constructing such a multigrid method. Since we use a conforming geometric multigrid method, the coarse-grid correction can be chosen canonically. Therefore, the smoother is actually the only degree of freedom in constructing the method. In this framework the smoother will be constructed such that the multigrid convergence rates are independent of the grid level. Obtaining multigrid convergence rates independent of the grid level is a main reason for choosing a multigrid method, as many other iterative methods do not allow this. The optimal control problems of our interest depend on a parameter and we are interested in the construction of multigrid solvers that allow also robustness of the convergence rates in this parameter, which is one challenge of this work. We will introduce two classes of smoothers: normal equation smoothers and collective point smoothers. Both classes of smoothers are used in practice and are known from literature. The first kind of smoothers, normal equation smoothers, are known to be relatively easy to analyze and have been proposed in literature, see, e.g., Brenner [22]. We will see in numerical

tests that the efficiency of those smoothers is comparable to other smoothers used in practice. The other choice, collective point smoothers, belongs to the class of Vanka smoothers, see, e.g., Vanka [67], and can be used without further knowledge for various problems. Such kind of smoothers have been proposed for optimal control problems, e.g., in Borzi, Kunisch and Kwak [12].

In this thesis, we will discuss mainly two kinds of convergence analysis. On the one hand, we stick to rigorous convergence proofs, based on a multiplicative splitting into smoothing property and approximation property, as introduced by Hackbusch, see, e.g., his book on multigrid, Hackbusch [35]. Already Brenner [22] introduced a framework for showing the convergence of a multigrid method for parameter-dependent saddle point problems satisfying certain properties. Unfortunately, her results cannot be directly applied to all model problems we consider in this thesis. We will give another convergence framework which follows another strategy. We will introduce five sufficient conditions for convergence of a multigrid method. The proof itself follows standard proofs for two-grid and W-cycle multigrid methods, which can be found in literature, e.g., in Hackbusch [35]. The framework covers on the one hand the approximation property and on the other hand the smoothing property for the smoothers based on the normal equation. The combination of both results implies convergence.

We apply this framework to three model problems to obtain convergence results for all of them. This extends the work of Simon [57] and Simon and Zulehner [58] to more general control problems. The extension of their work to the boundary control model problem was published in Takacs and Zulehner [61]. Afterwards, we will see how to extend the convergence analysis to obtain parameter-robust results. Such a result was already stated – within a different framework – in Schöberl, Simon and Zulehner [53]. In Takacs and Zulehner [62], we have extended their result to collective point smoothers, where rigorous analysis has not been available. Afterwards, we will relax the regularity assumptions which were necessary in Schöberl, Simon and Zulehner [53] and in the other papers cited above to the case of partial regularity (which allows to cover domains with reentrant corners).

The other approach for a multigrid convergence analysis, which we also study here, is local Fourier analysis (or local mode analysis), cf. Brandt [19]. The main idea is to use Fourier series in the analysis of multigrid methods. Local Fourier analysis provides a framework to analyze various numerical methods with a unified approach that gives quantitative statements on the methods under investigation and leads to the determination of sharp convergence rates. Local Fourier analysis can be justified rigorously only in special cases, e.g., on rectangular domains with uniform grids and periodic boundary conditions. However, results obtained with local Fourier analysis can be carried over to more general problems at least in a heuristic way.

The use of the Fourier series reduces the need of analyzing (discretized) differential operators and multigrid methods to the need of analyzing algebraic relations. For example, for computing the convergence rate of a multigrid method we will see that we have to determine the supremum of a rational function. So far, these algebraic problems have been solved by numerical approximation, see, e.g., the work Borzi, Kunisch and Kwak [12] on a local mode analysis for a model problem discussed also in this thesis. In a joint work with Veronika Pillwein, we have used quantifier elimination algorithms based on cylindrical algebraic decomposition, see, e.g., Collins [27], for computing such relations in an exact way. The results presented in this theses were published in Pillwein and Takacs [48].

This thesis is organized as follows. In Chapter 2 we will introduce the main framework. First, we will give some standard statements on Sobolev spaces in Section 2.1. In Section 2.2, we will present the optimal control model problems which will be discussed throughout the whole thesis. In Section 2.3 we will shortly discuss the concept of weak formulations and standard finite element methods that can be used to discretize the problem of our interest, which will lead to a linear system to be solved.

In Chapter 3 we will present the all-at-once multigrid methods which we propose for solving such systems. In Section 3.1 we will introduce the overall framework and comment on the coarse-grid correction. Subsequently, in Section 3.2 two classes of smoothers will be proposed: normal equation smoothers and collective point smoothers.

The analysis is done in two chapters. In Chapter 4 we will discuss the analysis based on Hackbusch's splitting into smoothing property and approximation property. At first we will introduce this splitting and we will derive general sufficient conditions for convergence of the proposed multigrid methods in Section 4.1. Here, we will follow proofs found in literature. In Section 4.2 we will apply this framework to the model problems to show the approximation property. Moreover, this section also cover the smoothing property for the normal equation smoothers, which together with the approximation property implies convergence. In Section 4.3, we will extend the convergence analysis to obtain parameter-robust results. Such a result was already stated – within a different framework – in Schöberl, Simon and Zulehner [53]. We will extend their result in Section 4.4 to collective point smoothers. In Section 4.5, we will relax the regularity assumptions introduced in Section 4.3 to the case of partial regularity.

In Chapter 5 we will discuss local Fourier analysis. Therefore, we will first introduce the local Fourier analysis framework in Section 5.1. Then we will discuss quantifier elimination, mention a method to perform quantifier elimination for a given formula and discuss its link to local Fourier analysis in Section 5.2. In Sections 5.3 and 5.4, we

will apply the framework to the model problem to derive sharp convergence results for the model problems.

In Chapter 6 we will present numerical results and in Chapter 7 we will give some concluding remarks.

Chapter 2

Preliminaries

In this chapter, we will give some definitions and preliminary results. We will start with Sobolev spaces, which are the natural spaces for considering the weak formulation of partial differential equations (PDEs).

In a next step, we will introduce the optimal control problems, which will be discussed throughout this thesis. Then we will discuss existence and uniqueness of the solution of the problem and the discretization of the problem. We will start with the partial differential equation forming the constraint, which is an elliptic equation. Afterwards, we will discuss existence and uniqueness of the solution of the optimization problem, the introduction of the optimality system and its discretization.

After the third section of this chapter, we will have a linear system to be solved. Then, in the last section of this chapter, we will discuss various iterative solvers for solving such a linear system.

2.1 Banach, Hilbert, Lebesgue and Sobolev spaces

In this section, we introduce the definitions and ideas related to Banach and Hilbert spaces, Lebesgue and Sobolev spaces. This is done to keep this thesis self-contained. Results are only presented if they are needed in later chapters of this thesis. Therefore, we restrict ourselves to the case of real vector spaces. For more details we refer to standard literature, e.g., Adams and Fournier [1] or Brenner and Scott [21].

A *Banach space* is a vector space A together with a norm $\|\cdot\|_A$, also written as $(A, \|\cdot\|_A)$ such that A is complete with respect to the norm $\|\cdot\|_A$, i.e., all Cauchy sequences are convergent. If the norm is induced by a scalar product, i.e., if

$$\|u\|_A = (u,u)_A^{1/2}$$

holds for all $u \in A$ and some scalar product $(\cdot,\cdot)_A : A \times A \to \mathbb{R}$, we call $(A,(\cdot,\cdot)_A)$ a *Hilbert space*.

The norm of the Banach space $(A, \|\cdot\|_A)$ is induced by a scalar product if and only if the parallelogram identity

$$\|u+v\|_A^2 + \|u-v\|_A^2 = 2\left(\|u\|_A^2 + \|v\|_A^2\right)$$

is satisfied for all u and $v \in A$. If this identity holds on A, the scalar product is given by

$$(u,v)_A := \frac{1}{4}\left(\|u+v\|_A^2 - \|u-v\|_A^2\right).$$

So, $(A,(\cdot,\cdot)_A)$ is Hilbert space. Due to the fact that the scalar product is characterized by the norm, we call also $(A, \|\cdot\|_A)$ a Hilbert space if the parallelogram identity is satisfied.

The set of bounded linear functionals mapping from a (reflexive) Banach space (Hilbert space) A to the reals $\mathbb{R}$ is again a Banach space (Hilbert space) equipped with norm

$$\|u\|_{A^*} := \sup_{v \in A\setminus\{0\}} \frac{u(v)}{\|v\|_A}$$

for all such functionals u. We call this set the dual of A, in short A^*. Often we use for the evaluation of such a functional the duality product $\langle u,v\rangle := u(v)$.

Having Banach (Hilbert) spaces, we can easily construct more such spaces by taking the algebraic sum or by taking the intersection. Let A_1 and A_2 be Banach (Hilbert) spaces with norms $\|\cdot\|_{A_1}$ and $\|\cdot\|_{A_2}$, respectively. The algebraic sum,

$$A_1 + A_2 = \{u_1 + u_2 \,:\, u_1 \in A_1 \text{ and } u_2 \in A_2\},$$

and the intersection,

$$A_1 \cap A_2,$$

are also Banach (Hilbert) spaces equipped with norms

$$\|u\|_{A_1+A_2} = \inf_{u=u_1+u_2,\, u_1\in A_1,\, u_2\in A_2} (\|u_1\|_{A_1}^2 + \|u_2\|_{A_2}^2)^{1/2} \text{ and}$$
$$\|u\|_{A_1\cap A_2} = (\|u\|_{A_1}^2 + \|u\|_{A_2}^2)^{1/2},$$

respectively, see, e.g., Proposition 3.2.1 in Butzer and Berens [26]. We have moreover

$$(A_1 + A_2)^* = A_1^* \cap A_2^* \text{ and } (A_1 \cap A_2)^* = A_1^* + A_2^*.$$

Now we can introduce the Lebesgue space L^2 and the Sobolev spaces H^m for $m \in \mathbb{N} := \{1, 2, 3, \ldots\}$. Here and in what follows, let Ω be a bounded open subset of $\mathbb{R}^d$ (for $d \in \{1, 2, 3\}$) with Lipschitz boundary $\partial\Omega$ (we call Ω a *domain*).

First we introduce the standard *Lebesgue space*: $L^2(\Omega)$ is the set of all (real-valued) square-integrable functions on Ω (integrability is understood in the sense of Lebesgue integrals). On this vector space we can introduce the scalar product

$$(u, v)_{L^2(\Omega)} := \int_\Omega u(x)v(x)\,\mathrm{d}x$$

and the corresponding norm $\|\cdot\|_{L^2(\Omega)} := (\cdot,\cdot)^{1/2}_{L^2(\Omega)}$. Using the convention that two L^2-functions u and v are equal if their values are equal almost everywhere, one can show that $(\cdot,\cdot)_{L^2(\Omega)}$ is a scalar product and $(L^2(\Omega), (\cdot,\cdot)_{L^2(\Omega)})$ is a Hilbert space.

Using the concept of *weak derivatives*, we can introduce Sobolev spaces. Let $\alpha = (\alpha_1, \ldots, \alpha_d) \in (\mathbb{N}_0)^d := \{0, 1, 2, 3, \ldots\}^d$ be a multi-index. The function $w \in L^2(\Omega)$ is called the α-th weak derivative of $u \in L^2(\Omega)$, for short $w = \mathrm{D}^\alpha u$, if

$$(w, v)_{L^2(\Omega)} = (-1)^{|\alpha|}(u, \mathrm{D}^\alpha v)_{L^2(\Omega)} \qquad \text{for all } v \in C_0^\infty(\Omega),$$

where $C_0^\infty(\Omega)$ denotes all C^∞-functions that have compact support in Ω. Here, $|\alpha| = |\alpha_1| + \ldots + |\alpha_d|$ and D^α, applied to a C^∞-function, is the differential operator

$$\mathrm{D}^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_1^{\alpha_d}}.$$

For $m \in \mathbb{N}_0$, the *Sobolev space* $H^m(\Omega)$ is the set of all functions $u \in L^2(\Omega)$ such that for all multi-indices α with $|\alpha| \le m$ the weak derivative $\mathrm{D}^\alpha u \in L^2(\Omega)$ exists, i.e.,

$$H^m(\Omega) := \left\{ u \in L^2(\Omega) \,:\, \mathrm{D}^\alpha u \in L^2(\Omega) \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \le m \right\}. \tag{2.1}$$

Together with the scalar product

$$(u, v)_{H^m(\Omega)} := \sum_{|\alpha| \le m} (\mathrm{D}^\alpha u, \mathrm{D}^\alpha v)_{L^2(\Omega)},$$

the Sobolev space $H^m(\Omega)$ is a Hilbert space.

This definition covers the case $m = 0$ (due to (2.1), we have $H^0(\Omega) = L^2(\Omega)$). Dual spaces of Sobolev spaces are interpreted as follows. The dual space of $L^2(\Omega)$, denoted by $(L^2(\Omega))^*$, can be identified with $L^2(\Omega)$ itself. For $m > 0$, we interpret the dual spaces of $H^m(\Omega)$, denoted by $(H^m(\Omega))^*$, as supersets of $L^2(\Omega)$. To do so, we need an embedding of $L^2(\Omega)$ into spaces $(H^m(\Omega))^*$: we associate to any $u \in L^2(\Omega)$ the functional

$$(u, \cdot)_{L^2(\Omega)} \in (H^m(\Omega))^*.$$

Note that not every $v \in (H^m(\Omega))^*$ can be expressed in this form, therefore $L^2(\Omega)$ and $(H^m(\Omega))^*$ are non-equal. The dual spaces could be introduced as Sobolev spaces with negative index m. We do not use this notion here, as sets like $H^{-1}(\Omega)$ include special assumptions on boundary conditions. In Section 4.5, we will generalize the notation of Sobolev spaces for non-integer indices.

2.2 Optimal control problems

The next step is the introduction of the model problems. As mentioned, we are interested in a particular class of PDE-constrained optimization problems: optimal control problems. Such problems have the following abstract setting. We consider some system, where its state can be described using the variable (state variable) $y \in Y$. This can be, for example, a heat distribution, a flow field or the pressure distribution. We assume that the state variable y satisfies a PDE of the form

$$\begin{aligned} Ly &= f(u), \\ By &= g(u), \end{aligned}$$

where L is a differential operator, B is a boundary operator and u is a variable that describes parameters that can be adjusted from outside of the system, like forces or heating sources applied from outside. We assume that we can adjust u, therefore we call u the control variable. f and g are given functionals.

If the PDE and appropriate boundary conditions (which may also depend on the control) are fixed, we assume that we are able to solve the system, i.e., we may compute for a given choice of the control variable u the state y.

The main point of optimal control problems is that we are interested in finding the best choice of the control u such that some cost functional J is minimized. Of course, this could be done using an outer iteration that uses a solver for the corresponding simulation problem in a black-box manner. Such methods, which guarantee that the constraints are satisfied for every iterate, are also called feasible path methods. We

are interested in a direct approach to solve such problems, so we are interested in conditions characterizing the solution of the optimization problem, which could be solved in a second step. For this purpose, we have to use information on the structure of the problem. Therefore, we prescribe a particular class of cost functional first.

Popular choices for the cost functional are tracking functionals. They consist on the one hand of the difference between state y and some desired state y_D, i.e., on

$$\|y - y_D\|,$$

where $\|\cdot\|$ measures something like the distance, i.e., it is typically a norm or a seminorm. Often, here the L^2-norm is considered, see, e.g., Lions [42]. We stick to this choice. Since that such a problem needs regularization, we assume also to have a regularization term. Therefore the cost functional may look as follows:

$$J(y, u) = \|y - y_D\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2,$$

where $\alpha > 0$ is a parameter. Depending on the considered model, α may be a cost parameter (for the costs that are related to applying the control) or a regularization parameter. Especially in the case that α is a regularization parameter, the choice of small values for α is of particular interest. So, we are interested in efficient finite element solvers for such optimal control problems, in particular in solvers where the convergence rates can be bounded from above by constants independent of the choice of α.

2.2.1 Model problems

In this thesis, we present the theory and the numerical results for some model problems. In the later chapters, the convergence theory will not be restricted to the model problems only. Numerical results will be computed for the model problems, introduced in this section only.

All our model problems are elliptic optimal control problems, i.e., the PDE of our interest is elliptic. For sake of simplicity, we restrict to the Laplace-like PDE

$$-\Delta y + y = f.$$

In general, we can handle other elliptic PDEs in a similar way.

The first model problem is a distributed control problem, i.e., we assume to control the system using a source term living on the whole domain Ω or, more generally, on a subdomain $\Omega_2 \subseteq \Omega$. So, the PDE looks like

$$-\Delta y + y = \begin{cases} u & \text{in } \Omega_2 \\ 0 & \text{in } \Omega \backslash \Omega_2 \end{cases},$$

where $u \in L^2(\Omega_2)$ or, shorter, as

$$-\Delta y + y = E_\Omega u \text{ in } \Omega,$$

where

$$E_\Omega u := \begin{cases} u & \text{in } \Omega_2 \\ 0 & \text{in } \Omega \backslash \Omega_2 \end{cases},$$

i.e., E_Ω is an extension operator $L^2(\Omega_2) \to L^2(\Omega)$ which extends u with 0 outside of Ω_2. Of course, we additionally need boundary conditions. For the model problems we assume for sake of simplicity to have homogeneous Neumann boundary conditions, i.e.,

$$\frac{\partial y}{\partial n} = 0 \text{ on } \partial\Omega,$$

where $\frac{\partial}{\partial n}$ is the outer-normal derivative.

We choose the tracking functional consisting of L^2-norms but the tracking functional may just live on subset a $\Omega_1 \subseteq \Omega$. So, the first model problem looks as follows.

Model Problem 1 *Find the control $u \in L^2(\Omega_2)$ and the state $y \in H^1(\Omega)$ such that they minimize the cost functional J, given by*

$$J(y,u) = \frac{1}{2}\|y - y_D\|^2_{L^2(\Omega_1)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega_2)},$$

subject to the elliptic boundary value problem

$$-\Delta y + y = E_\Omega u \text{ in } \Omega \qquad \text{and} \qquad \frac{\partial y}{\partial n} = 0 \text{ on } \partial\Omega. \tag{2.2}$$

Especially, if we consider the theory, the fact that the norms above only live in parts of Ω causes troubles. Therefore, we introduce an easier model problem, where $\Omega_1 = \Omega_2 = \Omega$ is satisfied.

Model Problem 2 *Find the control $u \in L^2(\Omega)$ and the state $y \in H^1(\Omega)$ such that they minimize the cost functional J, given by*

$$J(y,u) = \frac{1}{2}\|y - y_D\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)},$$

subject to the elliptic boundary value problem

$$-\Delta y + y = u \text{ in } \Omega \qquad \text{and} \qquad \frac{\partial y}{\partial n} = 0 \text{ on } \partial\Omega. \tag{2.3}$$

The third model problem we consider, is similar to the first two model problems. It is a boundary control problem. Here, the control variable does not live in the interior of the object but on its boundary and affects the boundary conditions of the state y, not the right-hand side of the partial differential equation.

Model Problem 3 *Find the control $u \in L^2(\partial\Omega)$ and the state $y \in H^1(\Omega)$ such that they minimize the cost functional J, given by*

$$J(y,u) = \frac{1}{2}\|y - y_D\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\partial\Omega)},$$

subject to the elliptic boundary value problem

$$-\Delta y + y = 0 \text{ in } \Omega \qquad \text{and} \qquad \frac{\partial y}{\partial n} = u \text{ on } \partial\Omega.$$

2.3 Weak formulations and discretization

In this section, we will introduce standard finite element techniques for discretizing the problem of our interest. The first part here is the introduction of the weak formulation and the discussion of existence and uniqueness. The second part consists of a finite element discretization. We will start both parts with the state equation only, as this equation itself is a standard elliptic model problem, which allows to introduce the standard elliptic convergence theory. Afterwards, we will extend the ideas in each case to saddle point problems, in particular, to the optimality systems associated with optimal control problems.

2.3.1 Weak formulation of the state equation

First we consider the state equation (2.2) in weak formulation, which reads as follows. Find $y \in Y := H^1(\Omega)$ such that

$$(\nabla y, \nabla \tilde{y})_{L^2(\Omega)} + (y, \tilde{y})_{L^2(\Omega)} = (u, \tilde{y})_{L^2(\Omega_2)} \qquad \text{for all } \tilde{y} \in Y.$$

Using the bilinear form b, given by

$$b(x, \tilde{y}) := (y, \tilde{y})_{H^1(\Omega)} = (\nabla y, \nabla \tilde{y})_{L^2(\Omega)} + (y, \tilde{y})_{L^2(\Omega)},$$

and the linear functional f,

$$f(\tilde{y}) := (u, \tilde{y})_{L^2(\Omega_2)},$$

we can rewrite the weak formulation as follows: Find $y \in Y := H^1(\Omega)$ such that

$$b(y, \tilde{y}) = f(\tilde{y}) \qquad \text{for all } \tilde{y} \in Y. \tag{2.4}$$

The first question to be answered is the question of existence and uniqueness of a solution y for a given control u. For elliptic problems, the Lax-Milgram theorem can be used to show existence and uniqueness.

Theorem 4 (Lax and Milgram) *Let $(Y, (\cdot,\cdot)_Y)$ be a Hilbert space. Let $b : Y \times Y \to \mathbb{R}$ be a bilinear form, which is*

- bounded, *i.e., there is a constant $\overline{C}$ such that*

$$b(y, \tilde{y}) \leq \overline{C}^2 \|y\|_Y \|\tilde{y}\|_Y \qquad \textit{for all } y,\, \tilde{y} \in Y \tag{2.5}$$

and

- coercive, *i.e., there is a constant $\underline{C}$ such that*

$$b(y, y) \geq \underline{C}^2 \|y\|_Y^2 \qquad \textit{for all } y \in Y. \tag{2.6}$$

Assume that $f \in Y^*$. Then, for a given f, there is exactly one solution y_f of the problem (2.4)*. The solution y_f satisfies*

$$\frac{1}{\overline{C}} \|f\|_{Y^*} \leq \|y_f\|_Y \leq \frac{1}{\underline{C}} \|f\|_{Y^*}.$$

For a proof see, e.g., Brenner and Scott [21], Theorem (2.7.7).

For the model problem the conditions of the Lax-Milgram theorem are satisfied with $\underline{C} = \overline{C} = 1$ because $b(\cdot,\cdot) = (\cdot,\cdot)_{H^1(\Omega)}$, i.e., the bilinear form is equal to the scalar product. Therefore existence of a solution and its uniqueness are guaranteed.

The conditions of the Lax-Milgram theorem imply the weaker condition

$$\underline{C}\|y\|_Y \leq \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{b(y,\tilde{y})}{\|\tilde{y}\|_Y} \leq \overline{C}\|y\|_Y \qquad \text{for all } y \in Y.$$

We will see in the next subsection that also this weaker condition is sufficient for showing existence and uniqueness. Here, we mention this condition because we will use this condition for showing existence and uniqueness of the solution of the optimal control model problems.

As proposed, we have shown that the state equation is solvable for every choice of the control variable $u \in L^2(\Omega_2)$ for Model Problem 1. The analysis can be extended to Model Problem 3 in a straight-forward way. So we can start discussing the introduction of the optimality systems.

2.3.2 Weak formulation of the control problem (Karush Kuhn Tucker system, KKT-system)

As mentioned above, we are interested in a characterization of the solution of the optimal control problem by a system of PDEs. Here, we use the method of Lagrange multipliers. For Model Problem 1, we obtain the Lagrange functional

$$\mathcal{L}(y,u,p) = \frac{1}{2}\|y - y_D\|^2_{L^2(\Omega_1)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega_2)} + (y,p)_{H^1(\Omega)} - (u,p)_{L^2(\Omega_2)}.$$

Solving the model problem is equivalent to finding a saddle point of the Lagrange functional which leads to the first order optimality system, which are called *Karush Kuhn Tucker system* (*KKT-system*). This system reads as follows. Find $(y,u,p) \in H^1(\Omega) \times L^2(\Omega_2) \times H^1(\Omega)$ such that

$$\begin{array}{lrlcl} (y,\tilde{y})_{L^2(\Omega_1)} & & + & (p,\tilde{y})_{H^1(\Omega)} & = & (y_D,\tilde{y})_{L^2(\Omega_1)} \\ & \alpha\,(u,\tilde{u})_{L^2(\Omega_2)} & - & (p,\tilde{u})_{L^2(\Omega_2)} & = & 0 \\ (y,\tilde{p})_{H^1(\Omega)} & - \quad (u,\tilde{p})_{L^2(\Omega_2)} & & & = & 0 \end{array}$$

holds for all $(\tilde{y},\tilde{u},\tilde{p}) \in H^1(\Omega)\times L^2(\Omega_2)\times H^1(\Omega)$. This system characterizes the solution of the Model Problem 1, cf. Tröltzsch [65] and others.

Because

$$\alpha\,(u,\tilde{u})_{L^2(\Omega_2)} = (p,\tilde{u})_{L^2(\Omega_2)}$$

holds for all $\tilde{u} \in L^2(\Omega_2)$, we obtain

$$u = \alpha^{-1} p|_{\Omega_2},$$

where $p|_{\Omega_2}$ is the restriction of the Lagrange multiplier p to the domain Ω_2. This allows us to reduce the KKT-system as follows. Find $(y,p) \in X := Y \times P := H^1(\Omega) \times H^1(\Omega)$ such that

$$\begin{array}{lcrcl} (y,\tilde{y})_{L^2(\Omega_1)} & + & (p,\tilde{y})_{H^1(\Omega)} & = & (y_D,\tilde{y})_{L^2(\Omega_1)} \\ (y,\tilde{p})_{H^1(\Omega)} & - & \alpha^{-1}(p,\tilde{p})_{L^2(\Omega_2)} & = & 0 \end{array}$$

holds for all $(\tilde{y},\tilde{p}) \in X$. This system is called the *reduced KKT-system.*

Obviously, this problem can also be interpreted as one single variational equation: Find $x \in X$ such that

$$\mathcal{B}(x,\tilde{x}) = \mathcal{F}(\tilde{x}) \qquad \text{for all } \tilde{x} \in X,$$

where

$$\begin{aligned} \mathcal{B}((y,p),(\tilde{y},\tilde{p})) &:= (y,\tilde{y})_{L^2(\Omega_1)} + (p,\tilde{y})_{H^1(\Omega)} + (y,\tilde{p})_{H^1(\Omega)} - \alpha^{-1}(p,\tilde{p})_{L^2(\Omega_2)}, \\ \mathcal{F}(\tilde{y},\tilde{p}) &:= (y_D,\tilde{y})_{L^2(\Omega_1)}. \end{aligned}$$

Now the question arises if this problem has a solution and if the solution is unique. As already mentioned, this question can be answered using the Babuška-Aziz Theorem.

Theorem 5 (Babuška and Aziz) *Let $(X_1, \|\cdot\|_{X_1})$ and $(X_2, \|\cdot\|_{X_2})$ be Hilbert spaces and let $\mathcal{B} : X_1 \times X_2 \to \mathbb{R}$ be a bilinear form and let $\mathcal{F} \in X_2^*$.*

Assume that there are constants $\underline{C} > 0$ and $\overline{C}$ such that

$$\underline{C}\|x\|_{X_1} \le \sup_{\tilde{x}\in X_2\setminus\{0\}} \frac{\mathcal{B}(x,\tilde{x})}{\|\tilde{x}\|_{X_2}} \le \overline{C}\|x\|_{X_1}$$

holds for all $x \in X_1$ and

$$\underline{C}\|x\|_{X_2} \le \sup_{\tilde{x}\in X_1\setminus\{0\}} \frac{\mathcal{B}(\tilde{x},x)}{\|\tilde{x}\|_{X_1}} \le \overline{C}\|x\|_{X_2}$$

holds for all $x \in X_2$.

Then the problem, find $x \in X_1$ such that

$$\mathcal{B}(x,\tilde{x}) = \mathcal{F}(\tilde{x}) \qquad \textit{for all } \tilde{x} \in X_2,$$

has a unique solution $x_{\mathcal{F}}$, which satisfies

$$\frac{1}{\overline{C}}\|\mathcal{F}\|_{X_2^*} \leq \|x_{\mathcal{F}}\|_{X_1} \leq \frac{1}{\underline{C}}\|\mathcal{F}\|_{X_2^*}.$$

For a proof see, e.g., Babuška [4], Theorem 2.1.

For the KKT-systems, the bilinear form $\mathcal{B}$ is symmetric. In the present section, we need the case $X_1 = X_2 := X$ and $\|\cdot\|_{X_1} = \|\cdot\|_{X_2} := \|\cdot\|_X$ only. Therefore, the two conditions of the Babuška-Aziz theorem (Theorem 5) reduce to the following condition:

(A1) There are constants $\underline{C} > 0$ and $\overline{C}$ such that the estimate

$$\underline{C}\|x\|_X \leq \sup_{\tilde{x}\in X\setminus\{0\}} \frac{\mathcal{B}(x,\tilde{x})}{\|\tilde{x}\|_X} \leq \overline{C}\|x\|_X$$

holds for all $x \in X$.

It is not easy to analyze **(A1)** directly for block-systems, like the (reduced) KKT-systems. As mentioned, these systems have saddle point structure. Therefore, we introduce a framework of saddle point problems and rephrase condition **(A1)** in terms of the individual blocks of the saddle point system.

Definition 6 (Saddle point problem) *We call the variational problem, find $x \in X$ such that*

$$\mathcal{B}(x,\tilde{x}) = \mathcal{F}(\tilde{x}) \qquad \text{for all } \tilde{x} \in X, \tag{2.7}$$

a saddle point problem*, if there are Hilbert spaces Y and P such that $X = Y \times P$ and there are bilinear forms a, b and c such that*

- $\mathcal{B}((y,p),(\tilde{y},\tilde{p})) = a(y,\tilde{y}) + b(y,\tilde{p}) + b(\tilde{y},p) - c(p,\tilde{p})$
- *a and c are symmetric, i.e., $a(y,\tilde{y}) = a(\tilde{y},y)$ and $c(p,\tilde{p}) = c(\tilde{p},p)$*
- *a and c are non-negative, i.e., $a(y,y) \geq 0$ and $c(p,p) \geq 0$.*

Note that every $\mathcal{F} \in X^* = Y^* \times P^*$ can be represented as

$$\mathcal{F}(\tilde{y},\tilde{p}) = \hat{f}(\tilde{y}) + \hat{g}(\tilde{y}), \tag{2.8}$$

where $\hat{f} \in Y^*$ and $\hat{g} \in P^*$.

Using this notation, the saddle point problem (2.7) can be rewritten in block-notation as follows. Find $y \in Y$ and $p \in P$ such that

$$\begin{array}{rcrcll} a(y,\tilde{y}) & + & b(\tilde{y},p) & = & \hat{f}(\tilde{y}) & \text{for all } \tilde{y} \in Y, \\ b(y,\tilde{p}) & - & c(p,\tilde{p}) & = & \hat{g}(\tilde{p}) & \text{for all } \tilde{p} \in P. \end{array}$$

In such a setting, i.e., if $X = Y \times P$, a norm on X can be constructed using the norms on Y and P by

$$\|(y,p)\|_X = \left(\|y\|_Y^2 + \|p\|_P^2\right)^{1/2}$$

for all $y \in Y$ and $p \in P$. Throughout this thesis, we restrict to this case.

Assuming that the bilinear form c vanishes, Brezzi's theorem gives sufficient conditions for existence and uniqueness of the solution of the problem. We give a variant of Brezzi's theorem stating that condition **(A1)** is satisfied.

Theorem 7 (Brezzi) *Assume that a saddle-point problem in the sense of Definition 6 with $c = 0$ satisfies the following conditions:*

- *There are constants $\underline{C}_1$ and $\overline{C}_1$ such that*

$$\underline{C}_1\|p\|_P \leq \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{b(\tilde{y},p)}{\|\tilde{y}\|_Y} \leq \overline{C}_1\|p\|_P \qquad \text{for all } p \in P.$$

- *There are constants $\underline{C}_2$ and $\overline{C}_2$ such that*

$$a(y,\tilde{y}) \leq \overline{C}_2\|y\|_Y\|\tilde{y}\|_Y \qquad \text{for all } y,\, \tilde{y} \in Y$$

  and

$$a(y,\tilde{y}) \geq \underline{C}_2\|y\|_Y^2$$

  for all $y \in \ker(B) := \{y \in Y : b(y,\tilde{p}) = 0 \text{ for all } \tilde{p} \in P\}$.

Then condition **(A1)** *is satisfied. The constants $\underline{C}$ and $\overline{C}$ in* **(A1)** *only depend on the constants $\underline{C}_1$, $\overline{C}_1$, $\underline{C}_2$ and $\overline{C}_2$, introduced above.*

For a proof, see Proposition 1.1 in Brezzi [23].

Since the bilinear form c does not vanish for the reduced KKT-system, we cannot apply Brezzi's theorem. Therefore we use another approach, which was introduced in Zulehner [71].

Theorem 8 *For saddle point problems in the sense of Definition 6,* **(A1)** *is equivalent to*

(A1') *There are constants* $\underline{C}_1 > 0$, $\underline{C}_2 > 0$, $\overline{C}_1$ *and* $\overline{C}_2$ *such that*

$$\underline{C}_1\|y\|_Y \leq \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{a(y,\tilde{y})}{\|\tilde{y}\|_Y} + \sup_{\tilde{p}\in P\setminus\{0\}} \frac{b(y,\tilde{p})}{\|\tilde{p}\|_P} \leq \overline{C}_1\|y\|_Y \text{ for all } y \in Y,$$
$$\underline{C}_2\|p\|_P \leq \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{b(\tilde{y},p)}{\|\tilde{y}\|_Y} + \sup_{\tilde{p}\in P\setminus\{0\}} \frac{c(p,\tilde{p})}{\|\tilde{p}\|_P} \leq \overline{C}_2\|p\|_P \text{ for all } p \in P.$$

The constants $\underline{C}$ *and* $\overline{C}$ *in* **(A1)** *only depend on the constants* $\underline{C}_1$, $\underline{C}_2$, $\overline{C}_1$ *and* $\overline{C}_2$ *in* **(A1')** *and vice versa.*

For a proof, see Zulehner [71], Theorems 2.2 and 2.3.

We return to the analysis of the reduced KKT-system of Model Problem 1. The KKT-system has saddle point structure (Definition 6), where obviously $a(\cdot,\cdot) = (\cdot,\cdot)_{L^2(\Omega_1)}$, $b(\cdot,\cdot) = (\cdot,\cdot)_{H^1(\Omega)}$, $c(\cdot,\cdot) = \alpha^{-1}(\cdot,\cdot)_{L^2(\Omega_2)}$, $\hat{f}(\cdot) = (y_D,\cdot)_{L^2(\Omega_1)}$ and $\hat{g} = 0$.

If applied to Model Problem 1, for standard norms condition **(A1')** reads as follows: There are constants $\underline{C}_1 > 0$, $\underline{C}_2 > 0$, $\overline{C}_1$ and $\overline{C}_2$ such that

$$\underline{C}_1\|y\|_{H^1(\Omega)} \leq \sup_{\tilde{y}\in H^1(\Omega)\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega_1)}}{\|\tilde{y}\|_{H^1(\Omega)}} + \sup_{\tilde{p}\in H^1(\Omega)\setminus\{0\}} \frac{(y,\tilde{p})_{H^1(\Omega)}}{\|\tilde{p}\|_{H^1(\Omega)}} \leq \overline{C}_1\|y\|_{H^1(\Omega)}$$

and

$$\underline{C}_2\|p\|_{H^1(\Omega)} \leq \sup_{\tilde{y}\in H^1(\Omega)\setminus\{0\}} \frac{(p,\tilde{y})_{H^1(\Omega)}}{\|\tilde{y}\|_{H^1(\Omega)}} + \sup_{\tilde{p}\in H^1(\Omega)\setminus\{0\}} \frac{\alpha^{-1}(p,\tilde{p})_{L^2(\Omega_2)}}{\|\tilde{p}\|_{H^1(\Omega)}} \leq \overline{C}_2\|p\|_{H^1(\Omega)}$$

for all $y, p \in H^1(\Omega)$.

First we analyze boundedness. Using the Cauchy-Schwarz inequality and the fact that the L^2-norms $\|\cdot\|_{L^2(\Omega_1)}$ and $\|\cdot\|_{L^2(\Omega_2)}$ are bounded by the $H^1(\Omega)$-norm, boundedness is guaranteed with constants $\overline{C}_1 = 2$ and $\overline{C}_2 = 2\alpha^{-1}$.

Now we analyze the lower bounds. First we recognize that we have

$$\sup_{\tilde{p}\in H^1(\Omega)\setminus\{0\}} \frac{(y,\tilde{p})_{H^1(\Omega)}}{\|\tilde{p}\|_{H^1(\Omega)}} = \|y\|_{H^1(\Omega)}.$$

Since

$$\sup_{\tilde{y}\in H^1(\Omega)\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega_1)}}{\|\tilde{y}\|_{H^1(\Omega)}} \geq 0,$$

we have the first inequality with constant $\underline{C}_1 = 1$. The same argument can be applied for the second inequality to obtain $\underline{C}_2 = 1$. This shows condition **(A1')**. Using Theorems 5 and 8, and the fact that Model Problem 2 is a special case of Model Problem 1, we conclude as follows.

Theorem 9 *Let $\alpha > 0$ be fixed. For Model Problems 1 and 2 condition* **(A1)** *is satisfied for the following choices of the norm:*

$$\|x\|_X := (\|y\|_Y^2 + \|p\|_P^2)^{1/2}, \tag{2.9}$$

where

$$\|y\|_Y := \|y\|_{H^1(\Omega)} \qquad \textit{and} \qquad \|p\|_P := \|p\|_{H^1(\Omega)}.$$

Remark 10 *For Model Problem 3, the bilinear form and the right-hand side read as follows:*

$$\begin{aligned}\mathcal{B}((y,p),(\tilde{y},\tilde{p})) &= (y,\tilde{y})_{L^2(\Omega)} + (p,\tilde{y})_{H^1(\Omega)} + (y,\tilde{p})_{H^1(\Omega)} - \alpha^{-1}(p,\tilde{p})_{L^2(\partial\Omega)},\\ \mathcal{F}(\tilde{y},\tilde{p}) &= (y_D,\tilde{y})_{L^2(\Omega)}.\end{aligned}$$

Also in this case, one can show that in the same way as above that condition **(A1)** *is satisfied for all fixed choices of $\alpha > 0$ and the norms chosen in Theorem 9.*

The same can be done, as long as the bilinear forms $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ are non-negative and bounded in $H^1(\Omega)$.

Remark 11 *For the non-reduced KKT-system, a similar analysis can be carried out. We give the results for Model Problem 1, the other model problems can be treated analogously.*

Here, we have to comment on the notation first. The non-reduced KKT-system consists of the primal variables y and u and of the dual variable p. Due to the fact that the framework above is formulated for one primal variable in Y and one dual variable P, we take $(y,u) \in Y$ to be the primal variable and $p \in P$ to be dual variable. Still, x consists of all variables, i.e., $x = (y,u,p) \in X = Y \times P$.

For the non-reduced KKT-systems, the bilinear forms a, b and c and the functionals $\hat{f}$ and $\hat{g}$, introduced in Definition 6, read as follows:

$$\begin{aligned} a((y,u),(\tilde{y},\tilde{u})) &= (y,\tilde{y})_{L^2(\Omega_1)} + \alpha(u,\tilde{u})_{L^2(\Omega_2)}, \\ b((y,u),\tilde{p}) &= (y,\tilde{p})_{H^1(\Omega)} - (y,\tilde{p})_{L^2(\Omega_2)}, \\ c(p,\tilde{p}) &= 0, \\ \hat{f}(\tilde{y},\tilde{u}) &= (y_D,\tilde{y})_{L^2(\Omega_1)}, \\ \hat{g}(\tilde{p}) &= 0. \end{aligned}$$

Theorems 5 and 8 can be applied and lead to the following result. For every fixed $\alpha > 0$, condition **(A1)** *is satisfied for the choice of the norm:*

$$\|x\|_X := (\|(y,u)\|_Y^2 + \|p\|_P^2)^{1/2},$$

where

$$\|(y,u)\|_Y := (\|y\|_{H^1(\Omega)}^2 + \|u\|_{L^2(\Omega_2)}^2)^{1/2} \qquad \textit{and} \qquad \|p\|_P := \|p\|_{H^1(\Omega)}.$$

We have seen that the constants in **(A1')** (and therefore also the constants in **(A1)**) depend on the choice of α. Therefore this approach does not allow to prove convergence results with constants independent of α, i.e., we will have to do some refinement of the analysis to obtain such a result.

For doing such a refined analysis, we have to restrict ourselves to the reduced KKT-system for Model Problem 2. Here, we have to find appropriate norms $\|\cdot\|_Y$ and $\|\cdot\|_P$ such that

$$\underline{C}_1\|y\|_Y \le \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega)}}{\|\tilde{y}\|_Y} + \sup_{\tilde{p}\in P\setminus\{0\}} \frac{(y,\tilde{p})_{H^1(\Omega)}}{\|\tilde{p}\|_P} \le \overline{C}_1\|y\|_Y \qquad \text{for all } y \in Y$$

and

$$\underline{C}_2\|p\|_P \le \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{(p,\tilde{y})_{H^1(\Omega)}}{\|\tilde{y}\|_Y} + \sup_{\tilde{p}\in P\setminus\{0\}} \frac{\alpha^{-1}(p,\tilde{p})_{L^2(\Omega)}}{\|\tilde{p}\|_P} \le \overline{C}_2\|p\|_P \qquad \text{for all } p \in P$$

is satisfied with constants independent of α.

The choice $Y = P$ with $\|p\|_P = \alpha^{-1/2}\|p\|_Y$ allows to reduce these two conditions to the condition

$$\underline{C}_1\|y\|_Y \le \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega)}}{\|\tilde{y}\|_Y} + \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{\alpha^{1/2}(y,\tilde{y})_{H^1(\Omega)}}{\|\tilde{y}\|_Y} \le \overline{C}_1\|y\|_Y \qquad \text{for all } y \in Y.$$

It is easy to see, that

$$\sup_{\tilde{y}\in Y\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega)}}{\|\tilde{y}\|_Y} + \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{\alpha^{1/2}(y,\tilde{y})_{H^1(\Omega)}}{\|\tilde{y}\|_Y} \geq \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega)} + \alpha^{1/2}(y,\tilde{y})_{H^1(\Omega)}}{\|\tilde{y}\|_Y}$$

holds for all $y \in Y$. On the right-hand side, we have a coercive bilinear form $(y,\tilde{y})_{L^2(\Omega)} + \alpha^{1/2}(y,\tilde{y})_{H^1(\Omega)}$. For this bilinear, we have

$$\sup_{\tilde{y}\in L^2(\Omega)\cap\alpha^{1/4}H^1(\Omega)\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega)} + \alpha^{1/2}(y,\tilde{y})_{H^1(\Omega)}}{\|\tilde{y}\|_{L^2(\Omega)\cap\alpha^{1/4}H^1(\Omega)}} = \|y\|_{L^2(\Omega)\cap\alpha^{1/4}H^1(\Omega)},$$

which shows the lower bound for $\underline{C}_1 = \underline{C}_2 = 1$. Cauchy-Schwarz inequality shows the upper bound with constants $\overline{C}_1 = \overline{C}_2 = 2$. Therefore, we conclude as follows.

Theorem 12 *For Model Problem 2, condition* **(A1)** *is satisfied with constants independent of the parameter α for the following choices of the norm:*

$$\|x\|_X := \left(\|y\|_Y^2 + \|p\|_P^2\right)^{1/2}, \tag{2.10}$$

where

$$\|y\|_Y := \|y\|_{L^2(\Omega)\cap\alpha^{1/4}H^1(\Omega)} = \left(\|y\|_{L^2(\Omega)} + \alpha^{1/2}\|y\|_{H^1(\Omega)}\right)^{1/2}$$

and

$$\|p\|_P := \|p\|_{\alpha^{-1/2}L^2(\Omega)\cap\alpha^{-1/4}H^1(\Omega)} = \left(\alpha^{-1}\|p\|_{L^2(\Omega)} + \alpha^{-1/2}\|p\|_{H^1(\Omega)}\right)^{1/2}.$$

This norm was introduced in Schöberl and Zulehner [54]. In Zulehner [71] this norm was derived in a straight-forward way using interpolation.

Remark 13 *A similar analysis can be carried out as long as $c(\cdot,\cdot) = \alpha^{-1}a(\cdot,\cdot)$ is satisfied for some $\alpha > 0$. This covers the case of Model Problem 1 for $\Omega_1 = \Omega_2 \subseteq \Omega$.*

Remark 14 *For the non-reduced KKT-system, a similar analysis can be carried out. Again, we have to restrict ourselves to Model Problem 2. Condition* **(A1)** *is satisfied with constants robust in the parameter α for the following choices of the norm:*

$$\|x\|_X := \left(\|(y,u)\|_Y^2 + \|p\|_P^2\right)^{1/2},$$

where

$$\|(y,u)\|_Y := \left(\|y\|^2_{L^2(\Omega)\cap\alpha^{1/4}H^1(\Omega)} + \|u\|^2_{\alpha^{1/2}L^2(\Omega)}\right)^{1/2}$$

and

$$\|p\|_P := \|p\|_{\alpha^{-1/2}L^2(\Omega)\cap\alpha^{-1/4}H^1(\Omega)}.$$

2.3.3 Discretization

As in the last section, we first introduce the idea of finite element discretization for the state equation only. For sake of simplicity, we restrict ourselves to the two-dimensional case. Let $\Omega \subset \mathbb{R}^2$ be some polygonal domain (an open, connected set with polygonal boundary).

The discretization is done using standard techniques. We use a family of triangular meshes, which are identified by grid levels $k \in \mathbb{N}_0$. We assume that some coarsest triangular mesh (grid level $k = 0$) is prescribed. The grids on the grid levels $k \in \mathbb{N}$ are obtained by uniform refinement, i.e., we subdivide every element into four congruent elements, see Figure 2.1.

We assume that the meshes are admissible, i.e., we assume that the set of elements $(\delta_{k,i})_{i=1}^{T_0}$ satisfies:

- all elements $\delta_{k,i}$ are open triangles,
- they cover the whole domain, i.e., $\cup_{i=1}^{T_k} \overline{\delta_{k,i}} = \overline{\Omega}$, where $\overline{A}$ is the closure of A,
- they do not intersect, i.e., $\delta_{k,i} \cap \delta_{k,j} = \emptyset$ for $i \neq j$ and
- the intersection of the closures of two elements, i.e., $\overline{\delta_{k,i}} \cap \overline{\delta_{k,j}}$, is either the empty set, a common vertex, a common edge or the element itself.

One can show in a straight-forward way that, if an admissible set is uniformly refined, also the refined meshes are admissible.

Figure 2.1: Uniform refinement

For $k \in \mathbb{N}_0$, we denote the size of the largest edge of the mesh by h_k. Since we have uniform refinement, $h_k = 2^{-k} h_0$ holds.

Based on the subdivision of the domain into triangles, we can introduce on every grid level a set of discretized functions $Y_k \subseteq Y$. Also here we use the easiest choice: the Courant element. The functions are linear on each element $\delta_{k,i}$ and continuous on the whole domain, i.e.,

$$Y_k := \{y_k \in C^1(\overline{\Omega}) \,:\, y_k|_{\delta_{k,i}} \text{ is linear for all } i = 1, \dots T_k\}.$$

As we have to work with that set, we need a good characterization of the degrees of freedom or, in other words, we need a good basis for the set Y_k. Obviously, a linear function mapping $\mathbb{R}^2$ to $\mathbb{R}$ can be characterized by the values of the function at three points (assuming the points are not located on a straight line). Therefore the linear functions on the elements can be characterized by the values on the three vertices. Due to the fact that we require continuity, a value fixed for one vertex affects directly all elements sharing the same vertex. It is easy to see, that every piecewise linear function constructed by prescribing its values at the vertices of the triangular elements is continuous on the whole domain. So the overall number of degrees of freedom is the number of vertices (nodes), which is denoted by N_k.

Therefore prescribing the values on the nodes is equivalent to specifying a function in Y_k. We introduce a nodal basis $(\varphi_{k,i})_{i=1}^{N_k}$ for Y_k as follows. The basis functions $\varphi_{k,i}$ are elements of Y_k and they satisfy

$$\varphi_{k,i}(\mathbf{x_j}) = \begin{cases} 1 & \text{for} \quad i = j \\ 0 & \text{for} \quad i \neq j, \end{cases}$$

for all nodes $\mathbf{x_j}$. As mentioned above, this completely describes the function $\varphi_{k,i}$. Every function $y_k \in Y_k$ can be represented with respect to this basis, i.e.,

$$y_k = \sum_{i=1}^{N_k} y_{k,i} \varphi_{k,i},$$

where the coefficients $y_{k,i}$ form the coefficient vector $\underline{y}_k = (y_{k,i})_{i=1}^{N_k}$.

For the discretization of the problem we use Galerkin's principle and replace the original variational problem, find $y \in Y$ such that

$$b(y, \tilde{y}) = f(\tilde{y}) \qquad \text{for all } \tilde{y} \in Y,$$

by the problem, find $y_k \in Y_k$ such that

$$b(y_k, \tilde{y}_k) = f(\tilde{y}_k) \qquad \text{for all } \tilde{y}_k \in Y_k. \tag{2.11}$$

Using the nodal basis, we can introduce the stiffness matrix

$$B_k = (b(\varphi_{k,i}, \varphi_{k,j}))_{i,j=1}^{N_k}$$

and the right-hand-side vector

$$\underline{f}_k = (f(\varphi_{k,j}))_{j=1}^{N_k},$$

which allows to rewrite (2.11) in matrix-vector notation as follows. Find $\underline{y}_k \in \mathbb{R}^{N_k}$ such that

$$B_k \, \underline{y}_k = \underline{f}_k$$

holds.

Also for the discretized problem, we have to show existence and uniqueness of the solution, i.e., if the matrix B_k is non-singular. This can be shown again using the Lax-Milgram theorem (Theorem 4), due to the fact that all conditions of the Lax-Milgram theorem are satisfied also if applied to the vector space $Y_k \subseteq Y$. Moreover, we are interested in a discretization error estimate, which the following lemma provides.

Theorem 15 (Céa) *Let $(Y, (\cdot,\cdot)_Y)$ be a Hilbert space and let $Y_k \subseteq Y$ be a closed subspace of Y. Let b be a bilinear form mapping $Y \times Y \to \mathbb{R}$ and let $f \in Y^*$.*

Consider the original problem, find $y \in Y$ such that

$$b(y, \tilde{y}) = f(\tilde{y}) \qquad \textit{for all } \tilde{y} \in Y,$$

and the discretized problem, find $y_k \in Y_k$ such that

$$b(y_k, \tilde{y}_k) = f(\tilde{y}_k) \qquad \textit{for all } \tilde{y}_k \in Y_k.$$

Assume that conditions of the Lax-Milgram theorem (Theorem 4) are satisfied. Then both problems have a unique solution and moreover the following discretization error estimate holds:

$$\|y - y_k\|_Y \leq \frac{\overline{C}}{\underline{C}} \inf_{\tilde{y}_k \in Y_k} \|y - \tilde{y}_k\|_Y.$$

For a proof see, e.g., Brenner and Scott [21], Theorem (2.8.1).

The next step is to estimate the approximation error. Using an interpolation operator $\Pi_k : Y \to Y_k$, we can estimate the approximation error error by the interpolation error, i.e., we have

$$\inf_{\tilde{y}_k \in Y_k} \|y - \tilde{y}_k\|_Y \leq \|y - \Pi_k y\|_Y.$$

Certainly, this result is valid for all interpolation operators. We are interested in an interpolation operator such that $\|y - \Pi_k y\|_Y$ is small. The existence of such an operator states the following theorem.

Theorem 16 (Interpolation error estimate) *For a family of meshes obtained by uniform refinement, and a discretization using the Courant element, there is a constant C_I (independent of grid level) and on every grid level $k \in \mathbb{N}_0$ an interpolation operator $\Pi_k : H^1(\Omega) \to Y_k \subseteq H^1(\Omega)$ such that for all $0 \leq i \leq m \leq 2$ and for all $y \in H^m(\Omega)$*

$$\|y - \Pi_k y\|_{H^i(\Omega)} \leq C_I \, h_k^{m-i} \|y\|_{H^m(\Omega)}$$

is satisfied.

For a proof see, e.g., Brenner and Scott [21], Theorem (4.8.7) and Remark (4.8.11).

Remark 17 *If we are only interested in introducing the interpolation operator defined on $H^2(\Omega)$, we can define the interpolation operator $\Pi_k : H^2(\Omega) \to Y_k$ by defining $y_k := \Pi_k y$ to be that function in Y_k satisfying*

$$y_k(\boldsymbol{x_j}) = y(\boldsymbol{x_j}) \qquad \text{for all nodes } \boldsymbol{x_j}.$$

This interpolation operator is not well-defined $H^1(\Omega)$. See equation (4.8.2) in Brenner and Scott *[21] for an interpolation operator which is well-defined in $H^1(\Omega)$.*

Combining all these results, we obtain an overall approximation error result

$$\inf_{\tilde{y}_k \in Y_k} \|y - \tilde{y}_k\|_{H^1(\Omega)} \leq \frac{\overline{C}}{\underline{C}} C_I h_k \|y\|_{H^2(\Omega)}, \tag{2.12}$$

and the complete error analysis

$$\|y - y_k\|_{H^1(\Omega)} \leq \frac{\overline{C}}{\underline{C}} C_I h_k \|y\|_{H^2(\Omega)},$$

which allows to bound the error by the H^2-norm of the solution y.

Next we discuss an inequality that $\|y\|_{H^2(\Omega)}$ bounds from above by a constant times $\|f\|_{L^2(\Omega)}$. Such a result cannot be guaranteed in general, but on domains with smooth boundary (see, e.g., NECAS [44]) or on polygonal or polyhedral domains which are convex (see, e.g., DAUGE [29, 30]), the following regularity result can be guaranteed:

(R) *Full elliptic regularity:* There is a constant $C_R > 0$ such that the following result holds. For $f \in L^2(\Omega)$ let $y_f \in H^1(\Omega)$ be the solution of

$$(y_f, \tilde{y})_{H^1(\Omega)} = (f, \tilde{y})_{L^2(\Omega)} \qquad \text{for all } \tilde{y} \in H^1(\Omega).$$

Then $y_f \in H^2(\Omega)$ and

$$\|y_f\|_{H^2(\Omega)} \leq C_R \|f\|_{L^2(\Omega)}.$$

Remark 18 *Since y_f is assumed to be the solution of a homogeneous Neumann problem, we have moreover $y_f \in Y^+ := \left\{ y \in H^2(\Omega) \,:\, \frac{\partial y}{\partial n}\Big|_{\partial\Omega} = 0 \right\}$.*

Remark 19 *Based on the fact that $f = -\Delta y_f + y_f$ also the estimate*

$$\|f\|_{L^2(\Omega)} \leq \| -\Delta y_f + y_f\|_{L^2(\Omega)} \leq \|y\|_{H^2(\Omega)}$$

holds for all $y_f \in Y^+$.

Remark 20 *Due to the fact, that the sets $H^2(\Omega) \subseteq H^1(\Omega) \subseteq L^2(\Omega)$ are dense in each other, the regularity assumption* **(R)** *implies that*

$$\underline{C}_R \|y\|_{H^2(\Omega)} \leq \sup_{\tilde{y} \in H^1(\Omega)\setminus\{0\}} \frac{(y,\tilde{y})_{H^1(\Omega)}}{\|\tilde{y}\|_{L^2(\Omega)}} \leq \overline{C}_R \|y\|_{H^2(\Omega)} \text{ for all } y \in Y^+$$

and

$$\underline{C}_R \|y\|_{L^2(\Omega)} \leq \sup_{\tilde{y} \in Y^+\setminus\{0\}} \frac{(y,\tilde{y})_{H^1(\Omega)}}{\|\tilde{y}\|_{H^2(\Omega)}} \leq \overline{C}_R \|y\|_{L^2(\Omega)} \text{ for all } y \in H^1(\Omega)$$

for $\underline{C}_R = C_R^{-1}$ and $\overline{C}_R = 1$.

And vice versa, the Babuška-Aziz theorem (Theorem 5) shows that these two conditions imply regularity assumption **(R)** *with $C_R = \underline{C}_R^{-1}$.*

The regularity assumption **(R)** allows to state the following a-priori error estimate.

Corollary 21 *Under the assumptions of Cea's Lemma and provided regularity assumption* **(R)** *is satisfied, the discretization error can be bounded as follows*

$$\|y - y_k\|_{H^1(\Omega)} \leq \frac{\overline{C}}{\underline{C}} C_I C_R h_k \|f\|_{L^2(\Omega)}.$$

Proof: Combine (2.12) and **(R)**. □

In Section 4.5 we will discuss how the regularity assumptions can be relaxed; in such a case we typically do not obtain an a-priori bound that behaves like h_k, but like h_k^γ for some $\gamma \in (0, 1)$.

2.3.4 Discretization of saddle point problems (Mixed finite elements)

In this section, we discuss the discretization of KKT-systems. As mentioned in the last section, if the assumptions of the Lax-Milgram theorem are satisfied for a space Y, they are also satisfied for every subspace Y_k. Therefore, no matter which discretization is chosen, the discretized problem is solvable and the standard discretization error results (Céa's lemma) hold.

This is not the case for saddle point problems, condition **(A1)** does not imply the corresponding condition for the discretized problem, which reads as follows.

(A1a) There are constants $\underline{C}_D > 0$ and $\overline{C}_D$ such that on all grid levels k the estimate

$$\underline{C}_D \|x\|_X \leq \sup_{\tilde{x} \in X_k \setminus \{0\}} \frac{\mathcal{B}(x, \tilde{x})}{\|\tilde{x}\|_X} \leq \overline{C}_D \|x\|_X$$

holds for all $x \in X_k$.

The construction of a discretization that also satisfies condition **(A1a)** is not easy in general, see, e.g., the Stokes problem and many other problems with saddle-point structure. For the Stokes problem it is well-known that several discretization techniques, which work well for elliptic problems, do not satisfy condition **(A1a)** and lead to instability, see, e.g., Braess [14], Chapter III, § 6. Therefore, the discretization schemes have to be chosen carefully to satisfy the condition, like the Taylor-Hood element (cf. Brezzi and Fortin [24]) or the Crouzeix-Raviart element (cf. Crouzeix and Raviart [28]).

For the optimal control problems we consider, the choice of an appropriate discretization, is not a big deal. Again Theorem 8 can be used to show that **(A1a)** is equivalent to

(A1a') There are constants $\underline{C}_{D,1} > 0$, $\underline{C}_{D,2} > 0$, $\overline{C}_{D,1}$ and $\overline{C}_{D,2}$ such that

$$\underline{C}_{D,1}\|y\|_Y \leq \sup_{\tilde{y}\in Y_k\setminus\{0\}} \frac{a(y,\tilde{y})}{\|\tilde{y}\|_Y} + \sup_{\tilde{p}\in P_k\setminus\{0\}} \frac{b(y,\tilde{p})}{\|\tilde{p}\|_P} \leq \overline{C}_{D,1}\|y\|_Y \text{ for all } y \in Y_k,$$
$$\underline{C}_{D,2}\|p\|_P \leq \sup_{\tilde{y}\in Y_k\setminus\{0\}} \frac{b(\tilde{y},p)}{\|\tilde{y}\|_Y} + \sup_{\tilde{p}\in P_k\setminus\{0\}} \frac{c(p,\tilde{p})}{\|\tilde{p}\|_P} \leq \overline{C}_{D,2}\|p\|_P \text{ for all } p \in P_k.$$

Provided that $Y_k = P_k$, we can show **(A1a')** in the same way as we could show **(A1')** in Subsection 2.3.2.

The discretization of the model problems reads as follows. Instead of finding $x \in X = Y \times P$ such that

$$\mathcal{B}(x,\tilde{x}) = \mathcal{F}(\tilde{x}) \qquad \text{for all } \tilde{x} \in X,$$

we consider the following problem. Find $x_k \in X_k := Y_k \times P_k$ such that

$$\mathcal{B}(x_k,\tilde{x}_k) = \mathcal{F}(\tilde{x}_k) \qquad \text{for all } \tilde{x}_k \in X_k.$$

Here, the sets $Y_k = P_k$ are constructed as in the last subsection.

The second question that may arise concerns the *discretization error* estimate. For elliptic problem, bounds for the discretization error are stated by Céa's lemma (Theorem 15). Due to the fact, that we are interested in saddle point problems, we give a more general theorem which covers them.

Theorem 22 (Discretization error) *Let $(X,(\cdot,\cdot)_X)$ be a Hilbert space and let $X_k \subseteq X$ be a closed subspace of X. Let $\mathcal{B}$ be a symmetric bilinear form mapping $X \times X \to \mathbb{R}$ and let $\mathcal{F} \in X^*$.*

Consider the original problem, find $x \in X$ such that

$$\mathcal{B}(x,\tilde{x}) = \mathcal{F}(\tilde{x}) \qquad \textit{for all } \tilde{x} \in X,$$

and the discretized problem, find $x_k \in X_k$ such that

$$\mathcal{B}(x_k,\tilde{x}_k) = \mathcal{F}(\tilde{x}_k) \qquad \textit{for all } \tilde{x}_k \in X_k.$$

Assume that conditions **(A1)** *and* **(A1a)** *are satisfied. Then both problems have a unique solution and moreover the following discretization error estimate holds:*

$$\|x - x_k\|_X \leq \frac{\overline{C}}{\underline{C}} \inf_{\tilde{x}_k \in X_k} \|x - \tilde{x}_k\|_X.$$

For a proof see, e.g., Babuška [4], Theorem 2.2.

Here, we can again estimate the approximation error as it was done in the last subsection. Since we need a regularity result for the optimality systems of our interest, we postpone this discussion to Chapter 4.

As in the last subsection, we can use the nodal basis to rewrite the optimality system in matrix-vector notation as follows:

$$\underbrace{\begin{pmatrix} A_k & B_k \\ B_k & -C_k \end{pmatrix}}_{\mathcal{A}_k :=} \underbrace{\begin{pmatrix} \underline{y}_k \\ \underline{p}_k \end{pmatrix}}_{\underline{x}_k :=} = \underbrace{\begin{pmatrix} \underline{g}_k \\ 0 \end{pmatrix}}_{\underline{\hat{f}}_k :=} \tag{2.13}$$

with mass matrices A_k and C_k and stiffness matrix B_k, given by

$$\begin{aligned} A_k &= (a(\varphi_{k,i}, \varphi_{k,j}))_{i,j=1}^{N_k} & B_k &= (b(\varphi_{k,i}, \varphi_{k,j}))_{i,j=1}^{N_k} \\ C_k &= (c(\varphi_{k,i}, \varphi_{k,j}))_{i,j=1}^{N_k} & \underline{\hat{f}}_k &= (\hat{f}(\varphi_{k,i}))_{i=1}^{N_k}. \end{aligned}$$

Here and in what follows, an underlined quantity, like $\underline{x}_k$, denotes the coefficient vectors of the corresponding function, here x_k, with respect to the nodal basis chosen for the corresponding space, here X_k.

2.4 Iterative solvers

The problem of our interest (2.13) is a large-scale sparse linear system. Therefore it is of particular interest to use iterative solvers for constructing approximate solutions for such a system. In this section, we give a short overview on iterative solvers which are relevant for the problems of our interest.

As in the last section, we start our discussion in Subsection 2.4.1 with the symmetric and positive definite matrix B_k, representing the state equation. In a second step (Subsection 2.4.2) we will discuss iterative solvers for the saddle point system $\mathcal{A}_k \, \underline{x}_k = \underline{f}_k$, representing the whole optimal control problem.

2.4.1 Iterative solvers for symmetric positive definite problems

In this subsection we consider the following problem. Find $\underline{y}_k$ such that

$$B_k \, \underline{y}_k = \underline{f}_k \tag{2.14}$$

holds, where B_k is a symmetric and positive definite matrix and $\underline{f}_k$ is a given vector.

The easiest solution method for such a problem is the gradient method or Richardson's iteration, which is given by the following iteration formula. Given an initial guess $\underline{y}_k^{(0)}$, for $m = \mathbb{N}_0$, the iterates $\underline{y}_k^{(m)}$ are given by

$$\underline{y}_k^{(m+1)} = \underline{y}_k^{(m)} + \tau \left(\underline{f}_k - B_k \, \underline{y}_k^{(m)} \right).$$

Here, we choose a relaxation parameter $\tau > 0$. The convergence analysis of Richardson's iteration is rather simple. As B_k is symmetric, also the iteration matrix

$$\mathcal{M}_k := I - \tau B_k,$$

is symmetric. Therefore $\sigma(\mathcal{M}_k)$, the spectrum of $\mathcal{M}_k$, s given by

$$\sigma(\mathcal{M}_k) = 1 - \tau\sigma(B_k) \subset [1 - \tau\lambda_{\max}(B_k), 1 - \tau\lambda_{\min}(B_k)]$$

because $\sigma(B_k) \subset [\lambda_{\min}(B_k), \lambda_{\max}(B_k)]$. Here, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimal and the maximal eigenvalue of a matrix A, respectively. The method is convergent if and only if the spectral radius of $\mathcal{M}_k$ is smaller than 1. This is the case if and only if

$$0 < \lambda_{\min}(B_k) \leq \lambda_{\max}(B_k) < \frac{2}{\tau}. \tag{2.15}$$

We obtain the optimal convergence rate using

$$\tau^* := \frac{2}{\lambda_{\max}(B_k) + \lambda_{\min}(B_k)},$$

which leads to the convergence rate

$$q = \frac{\kappa(B_k) - 1}{\kappa(B_k) + 1},$$

where

$$\kappa(A) := \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

is the condition number of a matrix A.

This analysis shows two facts. On the one hand, Richardson's iteration only converges for positive definite problems, cf. (2.15). Therefore, the method is not directly applica-

ble to the KKT-systems, which have saddle point structure. Methods for saddle point problems will be discussed in the next subsection.

On the other hand we have seen that the convergence rate of Richardson's method depends on the condition number B_k. Because B_k is the discretization of a second order elliptic PDE, we have

$$\kappa(B_k) = \mathcal{O}(h_k^{-2}),$$

i.e., the asymptotic behavior of the condition number is like h_k^{-2} for k approaching infinity. Therefore, the convergence rate q approaches 1 for k approaching infinitely, in detail we obtain

$$q = 1 - \mathcal{O}(h_k^2).$$

Since a good approximation of the solution of the PDE requires sufficiently fine grids, methods, where the convergence rates do not depend on the grid size, are of particular interest. We observe that the original problem (2.14) is equivalent to the preconditioned problem

$$\hat{B}_k^{-1} B_k \, \underline{y}_k = \hat{B}_k^{-1} \, \underline{u}_k.$$

If $\hat{B}_k$ is a symmetric and non-singular matrix and if B_k is symmetric and positive definite, the above equation is self-adjoined with respect to the scalar product induced by the matrix B_k. If $\hat{B}_k$ is positive definite, the system above is also positive definite. The convergence rate in this case is bounded by

$$\frac{\kappa(\hat{B}_k^{-1} B_k) - 1}{\kappa(\hat{B}_k^{-1} B_k) + 1}.$$

Therefore, the linear operator $\hat{B}_k^{-1}$ shall be chosen such that on the one hand the condition number is small. On the other hand, it should be chosen such that $\hat{B}_k^{-1} \underline{p}_k$ can be computed efficiently for any vector $\underline{p}_k$. Note that an explicit representation of $\hat{B}_k$ or $\hat{B}_k^{-1}$ as matrix is not necessary.

One possibility to construct such preconditioners, is the idea of Schwarz type methods. The easiest choices for such preconditioners lead to the Jacobi method (which can be seen as an additive Schwarz method) and the Gauss-Seidel method (which can be seen as an multiplicative Schwarz method). Using the notation $\underline{y}_k^{(m)} = \left(y_i^{(m)}\right)_{i=1}^{N_k}$ and $B_k = (b_{ij})_{i,j=1}^{N_k}$, we can write *Jacobi iteration* as follows:

$$y_i^{(m+1)} := y_i^{(m)} + b_{ii}^{-1} \left(f_i - \sum_{j=1}^{N_k} b_{ij} \, y_j^{(m)} \right),$$

i.e., we compute the (local) residuals $r_i := \left(f_i - \sum_{j=1}^{N_k} b_{ij}\, y_j^{(m)}\right)$ on every point (node), solve the problem

$$b_{ii}\, w_i = r_i$$

and update the iterate, i.e., set $y_i^{(m+1)} := y_i^{(m)} + w_i$. In case of *Gauss-Seidel iteration*, the newly computed updates are already used for computing the next residuals, i.e., we obtain

$$y_i^{(m+1)} = y_i^{(m)} + b_{ii}^{-1} \left(f_i - \sum_{j=1}^{i-1} b_{ij}\, y_j^{(m+1)} - \sum_{j=i}^{N_k} b_{ij}\, y_j^{(m)} \right),$$

Both methods, Jacobi iteration and Gauss-Seidel iteration, can be rewritten as preconditioned Richardson method. In case of Jacobi iteration, the preconditioner $\hat{B}_k$ is the diagonal of B_k and in case of the Gauss-Seidel method it is the lower-triangular part of B_k (including the diagonal).

Jacobi method and Gauss-Seidel method do not improve the convergence behavior qualitatively as also the condition number of the preconditioned system increases like $\mathcal{O}(h_k^{-2})$, if k approaches infinity. One kind of methods that guarantees condition numbers independent of the grid size, are multigrid and multilevel methods, which are based on sequences of grids. We will discuss multigrid methods in detail in Chapter 3.

More efficient methods than simple linear iteration schemes are typically Krylov subspace methods. For problems with a symmetric and positive matrix, we can apply the conjugate gradient methods, see Hestenes and Siefel [38]. For this method we obtain the convergence rate

$$\frac{\sqrt{\kappa(B_k)} - 1}{\sqrt{\kappa(B_k)} + 1}.$$

The conjugate gradient method converges, if the system matrix B_k is self-adjoined and positive definite. Note that also for the conjugate gradient method, the convergence rates increase if k approaches infinity.

The convergence rates can be improved using a preconditioner. For the preconditioned version of the conjugate gradient method the convergence rate is given by

$$\frac{\sqrt{\kappa(\hat{B}_k^{-1} B_k)} - 1}{\sqrt{\kappa(\hat{B}_k^{-1} B_k)} + 1}.$$

Like in the case of Richardson's iteration, also in this case, we require that $\hat{B}_k^{-1} B_k$ is self-adjoined with respect to some scalar product. Since standard Jacobi iteration as well as multigrid and multilevel methods (if set up accordingly) can be represented as preconditioned Richardson iteration with a symmetric preconditioner $\hat{B}_k$, these methods can also be applied as preconditioner for conjugate gradient method. In case of

(accordingly chosen) multigrid and multilevel methods, convergence rates independent in the grid size h_k can be expected for the overall iteration.

2.4.2 Iterative solvers for saddle point problems

Here, we give a rough overview of possible iteration schemes for saddle point problems. For more details, the author refers to the survey paper Benzi, Golub and Liesen [8].

First we want to present two kinds of preconditioning techniques. For the first kind of preconditioning techniques, the preconditioned system is again self-adjoined and positive definite. In the second case, the preconditioned system is indefinite.

First we start with preconditioned normal equation solvers, which are applicable to general systems

$$\mathcal{A}_k \, \underline{x}_k = \underline{f}_k.$$

The first observation is that

$$\mathcal{A}_k^T \, \mathcal{A}_k \, \underline{x}_k = \mathcal{A}_k^T \, \underline{f}_k, \tag{2.16}$$

is equivalent to the original problem. The matrix $\mathcal{A}_k^T \, \mathcal{A}_k$ is symmetric and positive definite. Therefore methods, developed for symmetric and positive definite problems, can be applied.

The equation (2.16) can be interpreted as an optimality condition for the following optimization problem. Find x_k such that it minimizes the norm of the residual, i.e.,

$$\min_{x_k \in X_k} J(x_k), \qquad \text{where} \qquad J(x_k) := \left\| \mathcal{A}_k \, \underline{x}_k - \underline{f}_k \right\|_{\ell^2}^2$$

and $\| \cdot \|_{\ell^2}$ is the Euclidean norm. The functional can be represented by

$$J(x_k) = \sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\left(\mathcal{B}\left(x_k, \tilde{x}_k\right) - \mathcal{F}\left(\tilde{x}_k\right) \right)^2}{\| \underline{\tilde{x}}_k \|_{\ell^2}^2}.$$

Here, instead of $\| \cdot \|_{\ell^2}$, another norm can be chosen. Consider

$$\min_{x_k \in X_k} J_X(x_k), \qquad \text{where} \qquad J_X(x_k) := \sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\left(\mathcal{B}(x_k, \tilde{x}_k) - \mathcal{F}(\tilde{x}_k) \right)^2}{\| \tilde{x}_k \|_X^2}. \tag{2.17}$$

This can be rewritten in matrix-vector notation. Let the symmetric and positive definite matrix $\mathcal{Q}_k$ be such that

$$(\mathcal{Q}_k \, \underline{x}_k, \underline{\tilde{x}}_k)_{\ell^2} = (x_k, \tilde{x}_k)_X \qquad \text{for all } x_k, \tilde{x}_k \in X_k.$$

Then the minimization problem (2.17) reads as follows.

$$\min_{x_k \in X_k} J_X(x_k), \qquad \text{where} \qquad J_X(x_k) = \|\mathcal{A}_k \underline{x}_k - \underline{f}_k\|^2_{\mathcal{Q}_k^{-1}}.$$

Here, the optimality condition reads as follows. Find $\underline{x}_k$ such that

$$\mathcal{Q}_k^{-1} \mathcal{A}_k^T \mathcal{Q}_k^{-1} \mathcal{A}_k \underline{x}_k = \mathcal{Q}_k^{-1} \mathcal{A}_k^T \mathcal{Q}_k^{-1} \underline{f}_k. \tag{2.18}$$

The system matrix $\mathcal{Q}_k^{-1} \mathcal{A}_k^T \mathcal{Q}_k^{-1} \mathcal{A}_k$ is self-adjoined in the scalar product $(\cdot,\cdot)_{\mathcal{Q}_k}$ and positive definite. The eigenvalues are bounded away from 0 and from above by constants that only depend on the constants in condition **(A1a)** . So, if the norm $\|\cdot\|_X$ is chosen such that the constants are independent of the grid level, we obtain optimal complexity, i.e., the convergence rates are bounded away from 1 by a constant independent of the grid level k. In case the constants are independent of the choice of α, also the convergence rates are robust in α.

Because the system matrix is self-adjoined, we can apply Richardson's method to the problem (2.18), which leads to the iteration

$$\underline{x}^{(m+1)} := \underline{x}^{(m)} + \tau \mathcal{Q}_k^{-1} \mathcal{A}^T \mathcal{Q}_k^{-1} (\underline{x}_k - \mathcal{A}_k \underline{x}_k^{(m)}),$$

where $0 < \tau < 2\rho(\mathcal{Q}_k^{-1/2} \mathcal{A}^T \mathcal{Q}_k^{-1/2})$ and $\rho(\cdot)$ denotes the spectral radius. One can show that, if τ is chosen in an optimal way, the convergence rates can be bounded by a constant that only depends on the constants in **(A1a)**. Variants of that method can be constructed by applying Jacobi or Gauss-Seidel iteration.

As the normal equation (2.18) is self-adjoined and positive definite, we can also apply conjugate gradient method as solver.

Certainly, for realizing the proposed methods based on the normal equation, we have to solve problems of the form: find $\underline{w}_k$ such that

$$\mathcal{Q}_k \underline{w}_k = \underline{r}_k$$

for some residual $\underline{r}_k$. The matrix $\mathcal{Q}_k$ is block-diagonal, therefore linear systems involving the blocks have to be solved. These blocks are symmetric and positive definite and involve the stiffness matrix (as the norm $\|\cdot\|_X$ is a (scaled) H^1-norm).

We have for the choice of $\|\cdot\|_X$ as in (2.9)

$$K_k \underline{w}_k = \underline{r}_k \tag{2.19}$$

and for the choice (2.10)

$$(M_k + \alpha^{1/2} K_k)\, \underline{w}_k = \underline{r}_k. \tag{2.20}$$

Here, methods presented in the last subsection can be used as the matrices K_k and $M_k + \alpha^{1/2} K_k$ are symmetric and positive definite. Since we have to solve four problems of the form (2.19) or (2.20) (at least approximately) for each iterate, also this may be costly.

More involved iteration schemes can be constructed by using structural information on the problem. For saddle point problems of the form

$$\mathcal{A}_k = \begin{pmatrix} A_k & B_k^T \\ B_k & -C_k \end{pmatrix},$$

iteration procedures based on a block LU-factorization have been proposed by Arrow, Hurwicz and Uzawa [3]. Bramble and Pasciak [18] proposed a possibility for preconditioning such a saddle point problem to obtain a self-adjoined and positive definite system also based on such a block LU-factorization. Assume that $\hat{A}_k$ is a symmetric and positive definite matrix with $\hat{A}_k > A_k$, i.e., such that $A_k - \hat{A}_k$ is positive definite. Then the preconditioned system $\hat{\mathcal{A}}_k^{-1} \mathcal{A}_k$ with

$$\hat{\mathcal{A}}_k = \begin{pmatrix} \hat{A}_k & 0 \\ B_k & -I \end{pmatrix},$$

is self-adjoined and positive definite with respect to the scalar product

$$\left(\left(\underline{y}_k, \underline{p}_k \right), \left(\tilde{\underline{y}}_k, \tilde{\underline{p}}_k \right) \right)_{BP} := \left(\left(A_k - \hat{A}_k \right) \underline{y}_k, \tilde{\underline{y}}_k \right)_{\ell^2} + \left(\underline{p}_k, \tilde{\underline{p}}_k \right)_{\ell^2}.$$

Therefore the conjugated gradient method can be applied (Bramble-Pasciak-CG, BPCG). Similar approaches were proposed, e.g., by Schöberl and Zulehner [54] or Benzi and Wathen [9].

Another possibility is to use Krylov subspace methods that can be directly applied to indefinite problems, e.g., the MINRES method, see Paige and Saunders [45]. This method can be applied directly to the discretized optimality system. For practical applications, good preconditioning is necessary. One possibility for computing such preconditioners are block-diagonal preconditioners. One possibility is again to use condition **(A1a)** as above, i.e., we use $\hat{\mathcal{A}}_k := \mathcal{Q}_k$ as block-diagonal preconditioner, see, e.g., Zulehner [71] for an application of that approach to the model problems discussed in this thesis. Another possibility is to choose

$$\hat{\mathcal{A}}_k := \begin{pmatrix} \hat{A}_k & \\ & \hat{S}_k \end{pmatrix},$$

where $\hat{A}_k$ approximates A_k and $\hat{S}_k$ approximates the Schur-complement $S_k = C_k - B_k A_k^{-1} B_k$. Such a method can be used to derive methods being robust in the parameter α for Model Problem 2, see, e.g., PEARSON AND WATHEN [47]. In both cases, two problems of the form (2.19) and (2.20) have to be solved for each iterate.

Another approach are all-at-once multigrid methods, which are directly applied to solve the whole saddle-point system. If such a method is introduced, the multigrid method can be used directly as a solver, i.e., an outer iteration scheme (like a MINRES method) is not necessary. Nonetheless, it would be also possible to use all-at-once multigrid methods for preconditioning a MINRES method.

We propose those all-at-once multigrid methods due to their flexibility and due to the fact that an outer iteration is not necessary. We will see that the construction of a multigrid iteration for the problem of our interest is non-standard, especially if the method should be constructed such that the convergence rates are robust in α.

Chapter 3

Multigrid methods

As mentioned in the last chapter, iterative solvers are essential tools for solving the linear system resulting from the discretization of partial differential equations. As we have seen in the last chapter, the convergence rates of (standard linear) iterative solvers typically depend on the condition number of the matrix. On the other hand, we have seen that the condition number of the stiffness matrix increases if the grid size h_k approaches 0.

Of course, if we want a higher accuracy of the approximate solution, it is important to refine the mesh accordingly. Therefore, methods, where the convergence rates do not depend on the grid size, are of particular interest. One class of methods having this property are multigrid methods. For an overview about multigrid methods, we refer to the books by Hackbusch [35], Bramble [17] and Trottenberg, Oosterlee and Schüller [66].

As the model problems of our interest are linear, we restrict ourselves to linear multigrid methods. A multigrid iteration scheme consists of two parts which have – in some sense – complementary properties: the *smoothing step* and and *coarse-grid correction step*. Intuitively speaking, the names smoothing and coarse-grid correction describe exactly what those two steps are doing.

A key observation for standard linear iteration schemes, like Jacobi iteration for symmetric and positive definite linear systems, reduce the hight-frequency parts of the residual rapidly. The fact that these methods have poor overall convergence rates is due to the fact, that low-frequency parts are reduced slowly.

The coarse-grid correction is based on restricting the residual to a coarser grid. The residual is typically approximated well on a coarse grid, if it consists of low-frequency parts only, as one can see for an easy one dimensional example in Figure 3.1. Of course,

much higher approximation errors have to be expected for oscillating functions, as one can see in Figure 3.2.

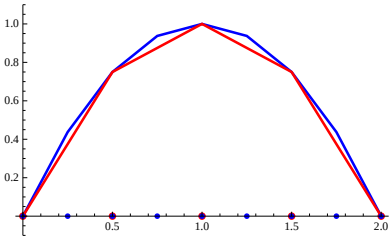


Figure 3.1: Approximation of a smooth function (blue) by a function on a coarser grid (red)

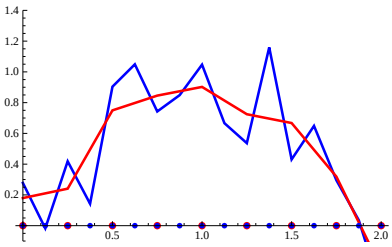


Figure 3.2: Approximation of an oscillating function (blue) by a function on a coarser grid (red)

Having these two iteration schemes with complementary properties, one could hope that the combination of these methods leads to a good iteration scheme.

This chapter is organized as follows. In Section 3.1 we will introduce the overall multigrid framework. Moreover we comment on the realization of the coarse-grid correction. We restrict ourselves to a canonical choice of the coarse-grid correction, which is possible in the framework we consider. For the construction of the smoother, there are several possibilities, which we will discuss in Section 3.2. In Section 3.3, we will discuss two possible strategies for convergence analysis. In Chapters 4 and 5, we will develop the convergence analysis for the model problems based on these strategies.

3.1 Multigrid framework

In this section, we introduce a general multigrid framework for solving linear systems arising from the discretization of a variational formulation which reads as follows. Find $x \in X$ such that

$$\mathcal{B}(x, \tilde{x}) = \mathcal{F}(\tilde{x}) \qquad \text{for all } \tilde{x} \in X.$$

Assume that we have a sequence of grids with grid levels $k \in \mathbb{N}_0$. On every grid level, we construct finite dimensional subsets $X_k \subseteq X$. Here, we assume that these subsets are nested, i.e., $X_k \subseteq X_{k+1}$ holds for all k. If the problem is discretized as as proposed in Subsection 2.3.3 using uniform refinement, we obtain nested subsets.

We have already mentioned the discretization in Chapter 2. Note that the problem can be discretized on each grid level and the discretized problems read as follows. Find $x_k \in X_k$ such that

$$\mathcal{B}(x_k, \tilde{x}_k) = \mathcal{F}(\tilde{x}_k) \qquad \text{for all } \tilde{x}_k \in X_k.$$

This can be rewritten in matrix-vector notation as follows. Find $\underline{x}_k \in \mathbb{R}^{N_k}$ such that

$$\mathcal{A}_k \, \underline{x}_k = \underline{f}_k. \tag{3.1}$$

The next step is the introduction of intergrid-transfer operators for the transfer between two consecutive grids. Since we consider nested subspaces, every function $x_{k-1} \in X_{k-1}$ is also an element of X_k. Therefore, the prolongation can be chosen in a canonical way: we choose the identity operator as prolongation operator. The matrix representation of the prolongation operator between X_{k-1} and X_k is denoted by $I_{k-1}^k \in \mathbb{R}^{N_{k-1} \times N_k}$.

The matrix representation I_k^{k+1} of the restriction operator is the transpose of the matrix representation of the prolongation operator, i.e., we choose $I_k^{k+1} := \left(I_{k-1}^k\right)^T$.

This allows to introduce the multigrid iteration for solving the discretized equation (3.1) on grid level k. Starting from an initial approximation $\underline{x}_k^{(0)}$, one step of the iteration is given in the following way:

- Apply ν *smoothing* steps:

$$\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m-1)} + \tau \, \hat{\mathcal{A}}_k^{-1} \left(\underline{f}_k - \mathcal{A}_k \, \underline{x}_k^{(0,m-1)} \right) \tag{3.2}$$

  for $m \in \{1, \ldots, \nu\}$ with $\underline{x}_k^{(0,0)} = \underline{x}_k^{(0)}$. The choice of τ and $\hat{\mathcal{A}}_k$ will be discussed below.

- Apply the *coarse-grid correction*, i.e.:

– Compute the defect $\underline{f}_k - \mathcal{A}_k \underline{x}_k^{(0,\nu)}$ and restrict it to the coarser grid (level $k-1$):

$$\underline{r}_{k-1}^{(1)} := I_k^{k-1} \left(\underline{f}_k - \mathcal{A}_k \underline{x}_k^{(0,\nu)} \right).$$

– Solve (approximatively) the linear system

$$\mathcal{A}_{k-1} \underline{w}_{k-1}^{(1)} = \underline{r}_{k-1}^{(1)}, \tag{3.3}$$

living on the coarser grid level $k-1$.

– Prolongate the result $\underline{w}_{k-1}^{(1)}$ to grid level k and add it to the last iterate:

$$\underline{x}_k^{(1)} := \underline{x}_k^{(0,\nu)} + I_{k-1}^k \underline{w}_{k-1}^{(1)}.$$

If the problem (3.3) is solved exactly, we obtain

$$\underline{x}_k^{(1)} = \underline{x}_k^{(0,\nu)} + I_{k-1}^k \mathcal{A}_{k-1}^{-1} I_k^{k-1} \left(\underline{f}_k - \mathcal{A}_k \underline{x}_k^{(0,\nu)} \right)$$

for the next iterate (two-grid method). In practice the solution of (3.3) is approximated by applying one step (V-cycle) or two steps (W-cycle) of the multigrid method, recursively. On grid level $k = 0$ the problem is solved exactly.

The idea of these multigrid iteration schemes is visualized in Figures 3.3 and 3.4, where the blue dots represent the smoothing steps, the red rectangles represent exact solves (on the coarsest grid) and the arrows represent the intergrid-transfer.

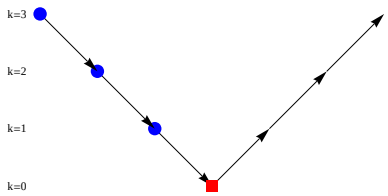


Figure 3.3: V-cycle multigrid method

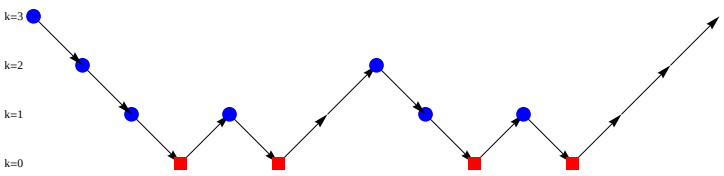


Figure 3.4: W-cycle multigrid method

Certainly, the multigrid method presented here can be modified in certain ways. We want to mention just one modification: the introduction of *post-smoothing*. The framework presented above consists of *pre-smoothing* steps and *coarse-grid correction*. After finishing coarse-grid correction, more smoothing steps (post-smoothing) can be applied.

For sake of simplicity, in Chapter 3, we restrict ourselves to pre-smoothing only and give remarks on the case that also post-smoothing is applied. In Chapters 4 and 5, we restrict ourselves to a method with pre- and post-smoothing.

3.2 Smoothers for saddle point problems

The next step in constructing a multigrid solver is the choice of the smoother.

For symmetric and positive definite systems arising from the discretization of partial differential equations, typically simple linear iteration schemes can be applied as smoothers. If we consider a finite element discretization of the Laplace equation, Richardson, Jacobi or Gauss-Seidel iteration are known to be good smoothers, see, e.g., HACKBUSCH [35], Chapter 3.3. Richardson and Jacobi iteration typically have to be damped if they are used as a smoother, whereas Gauss-Seidel iteration can by applied directly.

For saddle point problems, the choice of an appropriate smoother is a key issue. In this thesis we propose two classes of smoothers for saddle point problems, namely normal equation smoothers and collective smoothers.

3.2.1 Normal equation smoothers

In Subsection 2.4.1, we have already seen the approach of using the normal equation for constructing a solver. Here, we are not interested constructing an iteration scheme which is a good solver but in the construction of a good smoother.

In Subsection 2.4.1, the preconditioner was constructed based on the norm $\|\cdot\|_X$, represented by the matrix $\mathcal{Q}_k$.

Instead of $\|\cdot\|_X$, we use here another Hilbert space norm, $\|\cdot\|_{X_{-,k}}$, represented by a matrix $\mathcal{L}_k$. In Chapter 4, we will see how to choose this scalar product such that the overall multigrid method converges.

Here, we just assume that $\mathcal{L}_k$ is a symmetric and positive definite matrix.

Then the preconditioned normal equation smoother reads as follows:

$$\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m-1)} + \tau\, \mathcal{L}_k^{-1} \mathcal{A}_k^T \mathcal{L}_k^{-1} \left(\underline{f}_k - \mathcal{A}_k\, \underline{x}_k^{(0,m-1)} \right),$$

where τ is chosen such that

$$0 < \tau < \frac{2}{\rho\left(\mathcal{L}_k^{-1/2} \mathcal{A}_k \mathcal{L}_k^{-1/2} \right)^2}$$

holds. In principle, τ can be chosen on each grid level differently but we will see that, provided $\mathcal{L}_k$ is chosen accordingly, τ can be chosen to be equal on all grid levels.

We can also introduce variants of this smoother, especially in a Jacobi-type or in a Gauss-Seidel-type manner. The Jacobi-type normal equation smoother reads as follows:

$$\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m-1)} + \tau\, \mathrm{diag}\, \left(\mathcal{A}_k \mathcal{L}_k^{-1} \mathcal{A}_k \right)^{-1} \mathcal{A}_k \mathcal{L}_k^{-1} \left(\underline{f}_k - \mathcal{A}_k\, \underline{x}_k^{(0,m-1)} \right),$$

where τ is chosen such that

$$0 < \tau < \frac{2}{\rho\left(\mathrm{diag}\, \left(\mathcal{A}_k \mathcal{L}_k^{-1} \mathcal{A}_k \right)^{-1} \mathcal{A}_k \mathcal{L}_k^{-1} \mathcal{A}_k \right)}$$

holds. An advantage of the Jacobi-type normal equation smoother is the fact that is invariant with respect to a scaling of $\mathcal{L}_k$ by a scalar constant.

We will see in Chapter 4 how to choose $\mathcal{L}_k$ and τ in detail. Additionally, we should mention that all of these methods can be implemented in a reasonably efficient way if the matrix $\mathcal{L}_k$ is easily invertible.

3.2.2 Collective point smoothers

In this subsection we introduce the class of collective iteration schemes which relies on the block-structure of our problem. The iteration schemes are constructed by solving local problems, involving the complete system of PDEs, in an additive or multiplicative Schwarz-type manner. As in the case of elliptic problems, the local problems may live on patches or, as in our case, just on single points. Such methods have been proposed, e.g., in Trottenberg [66], Borzi, Kunisch and Kwak [12], Borzi and Schulz [13] and Lass [41].

For sake of simplicity, we focus on a method using local problems living on points only: *collective Jacobi iteration* for 2-by-2 block systems. Standard Jacobi relaxation, which can be used as a smoother for a linear system $B_k \underline{y}_k = \underline{f}_k$, where $B_k \in \mathbb{R}^{N_k \times N_k}$ is symmetric and positive definite, reads as follows:

$$y_i^{(0,m)} := y_i^{(0,m-1)} + \tau\, b_{ii}^{-1} \left(f_i - \sum_{j=1}^{N_k} b_{ij}\, y_j^{(0,m-1)} \right),$$

where $y_i^{(0,m)}$, f_i and b_{ij} are the components of the vectors $\underline{y}_k^{(0,m)}$ and $\underline{f}_k$ and the matrix B_k, respectively. This iteration scheme can be carried over to saddle point problems of the form

$$\underbrace{\begin{pmatrix} A_k & B_k \\ B_k & -C_k \end{pmatrix}}_{\mathcal{A}_k :=} \underbrace{\begin{pmatrix} \underline{y}_k \\ \underline{p}_k \end{pmatrix}}_{\underline{x}_k :=} = \underbrace{\begin{pmatrix} \hat{\underline{f}}_k \\ \hat{\underline{g}}_k \end{pmatrix}}_{\underline{f}_k :=}$$

in the following way. We define *collective Jacobi relaxation* to be the following iterative procedure:

$$\underline{x}_i^{(0,m)} := \underline{x}_i^{(0,m-1)} + \tau\, \mathcal{A}_{ii}^{-1} \left(\underline{f}_i - \sum_{j=1}^{N_k} \mathcal{A}_{ij}\, \underline{x}_j^{(0,m-1)} \right), \tag{3.4}$$

where $\underline{x}_i^{(0,m)} := \left(y_i^{(0,m)}, p_i^{(0,m)} \right)^T$, $\underline{f}_i := \left(\hat{f}_i, \hat{g}_i \right)^T$ and

$$\mathcal{A}_{ij} = \begin{pmatrix} a_{ij} & b_{ij} \\ b_{ij} & -c_{ij} \end{pmatrix}.$$

Here, $y_i^{(0,m)}$, $p_i^{(0,m)}$, $\hat{f}_i$, $\hat{g}_i$, a_{ij}, b_{ij} and c_{ij} are the components of $\underline{y}_k^{(0,m)}$, $\underline{p}_k^{(0,m)}$, $\hat{\underline{f}}_k$, $\hat{\underline{g}}_k$, A_k, B_k and C_k, respectively.

Collective Richardson relaxation and collective Gauss-Seidel relaxation are constructed analogously. Of course, such iteration schemes can be represented in a compact notation. The relaxation is given by the general formula (3.2) using the preconditioner

$$\hat{\mathcal{A}}_k = \begin{pmatrix} \hat{A}_k & \hat{B}_k \\ \hat{B}_k & -\hat{C}_k \end{pmatrix},$$

where $\hat{A}_k$, $\hat{B}_k$ and $\hat{C}_k$ are preconditioners for A_k, B_k and C_k, respectively. In particular:

- In the case of *collective Jacobi relaxation* $\hat{A}_k$, $\hat{B}_k$ and $\hat{C}_k$ are the diagonals of A_k, B_k and C_k, respectively, and the damping parameter τ is chosen to be in $(0, 1)$.

- In the case of *collective Richardson relaxation* we have $\hat{A}_k = a_k\, I$, $\hat{B}_k = b_k\, I$ and $\hat{C}_k = c_k\, I$, where for some constant $\hat{C} > 0$

$$\begin{aligned}
&\frac{1}{2}\lambda_{\max}(A_k) \leq a_k \leq \frac{\hat{C}}{2}\lambda_{\max}(A_k),\\
&\frac{1}{2}\lambda_{\max}(B_k) \leq b_k \leq \frac{\hat{C}}{2}\lambda_{\max}(B_k) \text{ and}\\
&\frac{1}{2}\lambda_{\max}(C_k) \leq c_k \leq \frac{\hat{C}}{2}\lambda_{\max}(C_k)
\end{aligned}$$

  holds. The damping parameter τ is chosen to be in $(0, 1)$.

- In the case of *collective Gauss-Seidel iteration* $\hat{A}_k$, $\hat{B}_k$ and $\hat{C}_k$ are the left-lower trigonal part (including the diagonal) of A_k, B_k and C_k, respectively, and the damping parameter τ is chosen to be 1.

Collective iteration schemes can be realized efficiently if they are implemented as described in (3.4), see e.g. Lass [41].

As we can easily see, collective iteration schemes can be introduced for the reduced KKT-systems for all three model problems. Contrary to the smoothers based on the normal equation, insight into the problem is not necessary for defining collective point smoothers.

3.2.3 Other classes of smoothers

Besides the classes of smoothers, we have introduced in the last two subsections, also other smoothers have been constructed for saddle point problems. Here, we want to mention only a few.

Uzawa type smoothers, that are based on a block LU-factorization of the iteration matrix based on an LU-factorization, have been applied in Simon and Zulehner [58] and Schöberl, Simon and Zulehner [53] to distributed control problems and in Takacs and Zulehner [61] to the boundary control Model Problem 3.

For problems with vanishing (2,2)-block, smoothers can be constructed such that the iterates stay in the subspace introduced by the (2,1)-block (constraint preconditioner). Braess and Sarazin [16] have proposed such a smoother with an simple (1,1)-block for the Stokes problem. One could consider such a smoother also for the non-reduced KKT-system.

Another class of smoothers are transforming smoothers, introduced in WITTUM [69, 70]. The idea of these smoothers is to transform the matrix $\mathcal{A}_k$ to a block-triangular form and to find smoothers for the diagonal blocks. This class of smoothers were applied to PDE-constrained problems in SCHULZ AND WITTUM [55].

3.3 Convergence analysis

We have seen that the proposed methods can be implemented in an efficient way and numerical experiments have shown that these methods work in practice. A main goal of this thesis is to confirm this observation by convergence theory. We have seen for the normal equation smoothers that the convergence analysis yields hints for the right choice of the matrix $\mathcal{L}_k$.

There are several ways of establishing convergence theory for multigrid methods. In this thesis, we consider two approaches: Hackbusch's multiplicative splitting into approximation property and smoothing property, which leads to rigorous convergence proofs. The other approach, we follow, is local Fourier analysis, which is based the fact that for simple grids (uniform, no boundaries) the error can be expressed in terms of Fourier series. This does not lead to a rigorous convergence proof for the general case but can be taken as an indicator for convergence of more general problems. Moreover, great advantages of local Fourier analysis are the facts that it is not only a qualitative analysis but also a quantitative analysis, i.e., it allows to compute sharp or at least realistic bounds for the convergence rate, and local Fourier analysis forms machinery which can be applied to various problems in a straight-forward way.

3.3.1 Smoothing and approximation property

To achieve a convergence result for the problems of our interest, we have to choose two norms, say $|||\cdot|||_{0,k}$ and $|||\cdot|||_{2,k}$. For convergence it is sufficient to show the following two conditions:

- *Smoothing property:*
  There is some function η with $\lim_{\nu\to\infty}\eta(\nu)=0$ such that for all grid levels $k\in\mathbb{N}$ and all $\nu\in\mathbb{N}$ the estimate

$$\left|\left|\left|\underline{x}_k^{(0,\nu)}-\underline{x}_k\right|\right|\right|_{2,k}\leq\eta(\nu)\,\left|\left|\left|\underline{x}_k^{(0)}-\underline{x}_k\right|\right|\right|_{0,k} \tag{3.5}$$

  holds.

- *Approximation property:* There is a constant $C_A > 0$ such that for all grid levels $k \in \mathbb{N}$ the estimate

$$\left|\left|\left| \underline{x}_k^{(1)} - \underline{x}_k \right|\right|\right|_{0,k} \leq C_A \left|\left|\left| \underline{x}_k^{(0,\nu)} - \underline{x}_k \right|\right|\right|_{2,k} \tag{3.6}$$

  holds.

Here, $\underline{x}_k := \mathcal{A}_k^{-1} \underline{f}_k$ is the exact solution of the linear system.

The combination of both estimates, (3.5) and (3.6), implies that the two-grid method converges if ν, the number of smoothing steps, is large enough. Due to standard arguments the convergence of the two-grid method implies the convergence of the W-cycle multigrid method under weak assumptions, as we will see in the next chapter. Hence analyzing smoothing and approximation property stated above, is of our particular interest.

This analysis is carried out in Chapter 4.

3.3.2 Local Fourier analysis

Another approach, which allows quantitative analysis, is local Fourier analysis. If we assume to have a regular grid and if we neglect the boundary conditions, i.e., we assume to have an infinite domain, we can compute the Fourier transformation of the system matrix $\mathcal{A}_k$ and of the components of the multigrid method.

The goal of such a transformation is the decoupling of the analysis for the individual Fourier modes which allows to determine exactly how the amplitude of each Fourier mode is modified by the multigrid iteration scheme. In case of the smoother, the analysis for the individual Fourier modes completely decouples. In case of the coarse-grid correction, linear spans of Fourier modes with small dimension have to be analyzed.

This analysis is carried out in Chapter 5. We will see that tools of symbolic computation can help to derive sharp bounds for the convergence rate.

Chapter 4

Multigrid analysis based on smoothing and approximation property

In this chapter we give convergence proofs for the all-at-once multigrid methods introduced in Chapter 3. The presentation of this chapter follows the framework and the notation introduced in the recent paper Takacs and Zulehner [63]. In Section 4.1, we will introduce a general convergence framework for multigrid methods, which follows classical ways. Then we will apply these conditions to the model problems in Sections 4.2 and 4.3.

As a part of the general convergence framework, we will also show that the normal equation smoother satisfies the smoothing property. Due to the fact that the convergence framework is constructed in a modular way, the proof of the approximation property can be combined with the proof of the smoothing property of any other smoother, provided that the norms used in the approximation property and the norms used in the smoothing property are equal. We will give examples of smoothing results that fit into the general framework. One of those examples, the class of collective point smoothers, will be worked out in detail in Section 4.4.

The results given in Sections 4.2 and 4.3 are based on the regularity assumption **(R)**, introduced on page 27. This regularity assumption cannot be guaranteed, e.g., for domains with reentrant corners. In Section 4.5, we will discuss the case that assumption **(R)** cannot be guaranteed, but a weaker condition **(R')**, which is satisfied also on non-convex polygonal domains.

4.1 A general convergence framework

The main goal of this section is the introduction of a systematic approach for the construction and the analysis of all-at-once multigrid methods for parameter-dependent saddle point problems. In this chapter we discuss the multigrid framework introduced in Section 3.1.

Already Brenner [22] introduced a framework for showing the convergence of a multigrid method for parameter-dependent saddle point problems satisfying certain properties. Unfortunately, her results cannot be applied directly to all model problems we consider in this thesis.

We will give another convergence framework which follows another strategy. We will introduce five sufficient conditions for convergence of a multigrid method. The proof itself follows standard proofs for two-grid and W-cycle multigrid methods, which can be found in literature, e.g., in Hackbusch [35]. The framework covers on the one hand the approximation property and on the other hand the smoothing property for the normal equation smoother. The combination of both results implies convergence.

For showing a multigrid convergence result based on Hackbusch's splitting of the analysis in smoothing property and approximation property, we have to introduce an appropriate framework. As mentioned, we choose for every grid level $k \in \mathbb{N}_0$ appropriate norms $|||\cdot|||_{0,k}$ and $|||\cdot|||_{2,k}$, defined on $\mathbb{R}^{N_k}$. Then it is sufficient to guarantee the smoothing property (3.5) and the approximation property (3.6), as introduced in Subsection 3.3.1.

For the framework, we consider, we need to be able to extend the norm $|||\cdot|||_{0,k}$ to an appropriate superset of X, i.e., we assume that there is a linear space $X_- \supseteq X$ equipped with (mesh-dependent) norms $\|\cdot\|_{X_{-,k}}$ and set

$$|||\underline{x}_k|||_{0,k} := \|x_k\|_{X_{-,k}}$$

for all $x_k \in X_k$ and all grid levels $k \in \mathbb{N}_0$. We assume that the norms $\|\cdot\|_{X_{-,k}}$ are induced by scalar products, therefore for all k, the tuples $X_{-,k} := (X_-, \|\cdot\|_{X_{-,k}})$ are Hilbert spaces.

Similar to other convergence frameworks, we choose $|||\cdot|||_{2,k}$ to be the residual norm with respect to $\|\cdot\|_{X_{-,k}}$, i.e., we have

$$|||\underline{x}_k|||_{2,k} := \sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\mathcal{B}(x_k, \tilde{x}_k)}{\|\tilde{x}_k\|_{X_{-,k}}}$$

for all $x_k \in X_k$ and all grid levels $k \in \mathbb{N}_0$.

For these choices, smoothing and approximation property read as follows.

- *Smoothing property:* There is some function η with $\lim_{\nu\to\infty} \eta(\nu) = 0$ such that for all grid levels $k \in \mathbb{N}$ and all $\nu \in \mathbb{N}$ the estimate

$$\sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\mathcal{B}\left(x_k^{(0,\nu)} - x_k, \tilde{x}_k\right)}{\|\tilde{x}_k\|_{X_{-,k}}} \leq \eta(\nu) \left\|x_k^{(0)} - x_k\right\|_{X_{-,k}} \tag{4.1}$$

  holds.

- *Approximation property:* There is a constant $C_A > 0$ such that for all grid levels $k \in \mathbb{N}$ the estimate

$$\left\|x_k^{(1)} - x_k\right\|_{X_{-,k}} \leq C_A \sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\mathcal{B}\left(x_k^{(0,\nu)} - x_k, \tilde{x}_k\right)}{\|\tilde{x}_k\|_{X_{-,k}}} \tag{4.2}$$

  holds.

It is easy to see that, if we combine both conditions, we obtain

$$\left\|x_k^{(1)} - x_k\right\|_{X_{-,k}} \leq q(\nu) \left\|x_k^{(0)} - x_k\right\|_{X_{-,k}},$$

where $q(\nu) = C_A\, \eta(\nu)$, i.e., the two-grid method converges if ν is sufficiently large. The convergence of the W-cycle multigrid method will be discussed in Subsection 4.1.4.

In the next two subsections, we will give five sufficient conditions that guarantee that the conditions (4.1) and (4.2) are satisfied. We will see that the analysis of those conditions is relatively easy and well known cases are covered by them.

4.1.1 Smoothing property for the normal equation smoother

In this subsection we give conditions on the choice of the norm $\|\cdot\|_{X_{-,k}}$ such that the normal equation smoother satisfies the smoothing property. As we have already mentioned, the knowledge on the norm $\|\cdot\|_{X_{-,k}}$ or more precisely its matrix representation $\mathcal{L}_k$, i.e.,

$$(\mathcal{L}_k\, \underline{x}_k, \underline{\tilde{x}}_k)_{\ell^2} = (x_k, \tilde{x}_k)_{X_{-,k}} \qquad \text{for all } x_k, \tilde{x}_k \in X_k.$$

is required for constructing the method. In this framework the smoothing property for the normal equation smoother can be shown:

Theorem 23 *Assume that condition* **(A1a)**, *introduced on page 28, and the following condition hold.*

(A2) *There is a constant $C_M > 0$ such that for all grid levels k the estimate*

$$\|x_k\|_X \leq C_M \|x_k\|_{X_{-,k}} \qquad \text{holds for all } x_k \in X_k.$$

Then for the preconditioned normal equation smoother

$$\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m-1)} + \tau\, \mathcal{L}_k^{-1} \mathcal{A}_k^T \mathcal{L}_k^{-1} \left(\underline{f}_k - \mathcal{A}_k\, \underline{x}_k^{(0,m-1)} \right), \tag{4.3}$$

with damping parameter $\tau := \tau(k) := \rho\left(\mathcal{L}_k^{-1/2} \mathcal{A}_k \mathcal{L}_k^{-1/2}\right)^{-2}$ satisfies the smoothing property with smoothing rate

$$\eta(\nu) = \frac{C_S}{\sqrt{\nu}},$$

where C_S only depends on the constants used in **(A1a)** *and* **(A2)**.

Proof: We have to show that there is a constant $C_S > 0$ such that for all choices of ν the following estimate holds:

$$\left\| \mathcal{L}_k^{-1/2} \mathcal{A}_k \mathcal{M}_k^\nu \mathcal{L}_k^{-1/2} \right\|_{\ell^2} \leq \frac{C_S}{\sqrt{\nu}},$$

where the iteration matrix $\mathcal{M}_k$ is given by

$$\mathcal{M}_k := I - \tau \mathcal{L}_k^{-1} \mathcal{A}_k^T \mathcal{L}_k^{-1} \mathcal{A}_k.$$

We obtain immediately

$$\begin{aligned}
\left\| \mathcal{L}_k^{-1/2} \mathcal{A}_k \mathcal{M}_k^\nu \mathcal{L}_k^{-1/2} \right\|_{\ell^2}^2 &= \left\| \mathcal{L}_k^{-1/2} \mathcal{A}_k (I - \tau \mathcal{L}_k^{-1} \mathcal{A}_k \mathcal{L}_k^{-1} \mathcal{A}_k)^\nu \mathcal{L}_k^{-1/2} \right\|_{\ell^2}^2 \\
&= \left\| \mathcal{L}_k^{-1/2} \mathcal{A}_k \mathcal{L}_k^{-1/2} (I - \tau \mathcal{L}_k^{-1/2} \mathcal{A}_k^T \mathcal{L}_k^{-1} \mathcal{A}_k \mathcal{L}_k^{-1/2})^\nu \right\|_{\ell^2}^2 \\
&= \left\| \mathcal{P}_k (I - \tau \mathcal{P}_k^T \mathcal{P}_k)^\nu \right\|_{\ell^2}^2
\end{aligned}$$

and using the definition of the Euclidean norm further

$$\begin{aligned}
\left\| \mathcal{L}_k^{-1/2} \mathcal{A}_k \mathcal{M}_k^\nu \mathcal{L}_k^{-1/2} \right\|_{\ell^2}^2 &= \rho\left((I - \tau \mathcal{P}_k^T \mathcal{P}_k)^\nu \mathcal{P}_k^T \mathcal{P}_k (I - \tau \mathcal{P}_k^T \mathcal{P}_k)^\nu \right) \\
&= \rho\left(\mathcal{P}_k^T \mathcal{P}_k (I - \tau \mathcal{P}_k^T \mathcal{P}_k)^{2\nu} \right),
\end{aligned} \tag{4.4}$$

where $\mathcal{P}_k := \mathcal{L}_k^{-1/2} \mathcal{A}_k \mathcal{L}_k^{-1/2}$. Certainly, $\mathcal{P}_k^T \mathcal{P}_k$ is symmetric and positive definite. Therefore all eigenvalues are positive and we can use an eigenvalue decomposition

to estimate (4.4). Using $\sigma := \left[\lambda_{min}\left(\mathcal{P}_k^T\mathcal{P}_k\right), \lambda_{max}\left(\mathcal{P}_k^T\mathcal{P}_k\right)\right]$ and the fact that $\tau \leq \lambda_{max}\left(\mathcal{P}_k^T\mathcal{P}_k\right)^{-1}$, we obtain

$$\begin{aligned}\rho\left(\mathcal{P}_k^T\mathcal{P}_k\left(I-\tau\mathcal{P}_k^T\mathcal{P}_k\right)^{2\nu}\right) &\leq \sup_{\lambda\in\sigma}\lambda(1-\tau\lambda)^{2\nu} \leq \sup_{\mu\in[0,1]}\tau^{-1}\mu(1-\mu)^{2\nu} \\ &= \tau^{-1}\frac{1}{(1+\nu^{-1})^{\nu}(1+\nu)} \leq \frac{1}{\tau\nu} \end{aligned} \tag{4.5}$$

Using conditions **(A1a)** and **(A2)**, we obtain

$$\begin{aligned}\tau^{-1} = \rho(\mathcal{P}_k)^2 &\leq \left\|\mathcal{L}_k^{-1/2}\mathcal{A}_k\mathcal{L}_k^{-1/2}\right\|_{\ell^2}^2 \\ &= \left\|\mathcal{L}_k^{-1/2}\mathcal{Q}_k^{1/2}\right\|_{\ell^2}^4 \left\|\mathcal{Q}_k^{-1/2}\mathcal{A}_k\mathcal{Q}_k^{-1/2}\right\|_{\ell^2}^2 \leq \overline{C}_D^2 C_M^4, \end{aligned} \tag{4.6}$$

where $\overline{C}_D$ is the constant from condition **(A1a)** and C_M is the constant from condition **(A2)**. By combining (4.5) and (4.6), we obtain

$$\left\|\mathcal{L}_k^{-1/2}\mathcal{A}_k\mathcal{M}_k^{\nu}\mathcal{L}_k^{-1/2}\right\|_{\ell^2}^2 \leq \overline{C}_D^2 C_M^4 \frac{1}{\nu} =: \frac{C_S^2}{\nu},$$

which finishes the proof. □

A second important property, which we will need in Subsection 4.1.3 and in Sections 4.4 and 4.5, is the concept of power-boundedness.

Lemma 24 *Under the assumptions of Theorem 23 there is a constant $C_B > 0$ such that for all grid levels k*

$$|||\mathcal{M}_k^{\nu}\,\underline{x}_k|||_{0,k} \leq C_B|||\underline{x}_k|||_{0,k} \tag{4.7}$$

holds and all $\underline{x}_k \in \mathbb{R}^{N_k}$ and all $\nu \in \mathbb{N}$. Here, $\mathcal{M}_k$ is the iteration matrix representing the normal equation smoother (4.3)*.*

Proof: The power-boundedness (4.7) can be rewritten in matrix-vector notation as follows

$$\left\|\mathcal{L}_k^{1/2}\mathcal{M}_k^{\nu}\mathcal{L}_k^{-1/2}\right\|_{\ell^2} \leq C_B. \tag{4.8}$$

Similar to the proof of Theorem 4.4, we can show that

$$\left\|\mathcal{L}_k^{1/2}\mathcal{M}_k\mathcal{L}_k^{-1/2}\right\|_{\ell^2} = \left\|I-\tau\mathcal{P}_k^T\mathcal{P}_k\right\|_{\ell^2} \leq 1,$$

which directly implies (4.8). □

In the next sections we will choose the norm $\|\cdot\|_{X_{-},k}$ such that the condition **(A2)**, introduced in Theorem 23, is satisfied. In the next subsection we will see that this norm has to satisfy also another condition to guarantee the approximation property.

For the realization of the smoother, we have to invert the matrix $\mathcal{L}_k$. Therefore we have to choose the norm $\|\cdot\|_{X_{-,k}}$ such that the matrix $\mathcal{L}_k$, representing this norm, (or a spectral equivalent matrix $\hat{\mathcal{L}}_k$) is easy to invert.

Note that condition **(A2)** is also of importance if other smoothers are considered because typically the smoothing property can only be shown if that condition is satisfied. The condition is also satisfied for multigrid methods proposed previously for the model problems in Schöberl, Simon and Zulehner [53], Takacs and Zulehner [62] and others.

Typically the *exact* implementation of the normal equation smoother is not efficient but the conditions can be relaxed further as follows.

Corollary 25 *The result stated in Theorem 23 also holds for an inexact version of the preconditioned normal equation smoother, where $\mathcal{L}_k$ is replaced by a spectrally equivalent matrix $\hat{\mathcal{L}}_k$, i.e., such that*

$$\underline{C}_{\mathcal{L}}\,\mathcal{L}_k \le \hat{\mathcal{L}}_k \le \overline{C}_{\mathcal{L}}\,\mathcal{L}_k,$$

holds for some constants $0 < \underline{C}_{\mathcal{L}} \le \overline{C}_{\mathcal{L}}$ and τ is replaced by $\hat{\tau}$ such that

$$\underline{C}_{\tau} \le \frac{\hat{\tau}}{\rho(\hat{\mathcal{P}}_k)^2} \le \overline{C}_{\tau},$$

holds for some constants $0 < \underline{C}_{\tau} \le \overline{C}_{\tau} < 2$, where $\hat{\mathcal{P}}_k := \hat{\mathcal{L}}_k^{-1/2}\mathcal{A}_k\hat{\mathcal{L}}_k^{-1/2}$.

In this case the smoothing rate may also depend on the constants $\underline{C}_{\tau}$, $\overline{C}_{\tau}$, $\underline{C}_{\mathcal{L}}$ and $\overline{C}_{\mathcal{L}}$.

We skip the proof of this corollary because the proof is a straight-forward but rather technical generalization of the proof of Theorem 23.

4.1.2 Approximation property

For showing the approximation property, first we reconsider known results for the Laplace equation. Consider the variational problem, find $y_k \in Y_k$ such that

$$b(y_k, \tilde{y}_k) = f(\tilde{y}_k) \qquad \text{for all } \tilde{y}_k \in Y_k,$$

which is the discretization of the original variational problem, find $y \in Y = H^1(\Omega)$ such that

$$b(y, \tilde{y}) = f(\tilde{y}) \qquad \text{for all } \tilde{y} \in Y.$$

If one can show boundedness (2.5) and coercivity (2.6) of the bilinear form b, the Lax-Milgram theorem (Theorem 4) guarantees existence and uniqueness of the solution. Cea's lemma (Theorem 15) allows to show the convergence error result

$$\|y - y_k\|_{H^1(\Omega)} \leq C\|f\|_{[H^1(\Omega)]^*}.$$

Using the approximation error result (2.12) and the regularity assumption **(R)**, introduced on page 27, we obtain

$$\|y - y_k\|_{H^1(\Omega)} \leq Ch_k\|f\|_{L^2(\Omega)},$$

Using a standard approach (Aubin Nitsche duality trick), one can show an estimate of the L^2-error:

$$\|y - y_k\|_{L^2(\Omega)} \leq Ch_k^2\|f\|_{L^2(\Omega)}. \tag{4.9}$$

Such a result is used to show the approximation property for multigrid methods in case of an elliptic problem, see, e.g., Hackbusch [35], Section 6.3.1.

Concluding, for a proof of the approximation property in case of an elliptic problem, we need that

1. the conditions of the Lax-Milgram theorem (estimates (2.5) and (2.6)),

2. the approximation error estimate (2.12) and

3. the regularity assumption **(R)**

are satisfied.

An analogous strategy can be applied for the indefinite problems of our interest, i.e., the ideas of Hackbusch's proof are not restricted to elliptic problems. We have already introduced the abstract Hilbert spaces X and $X_{-,k}$ (playing the role of $H^1(\Omega)$ and $L^2(\Omega)$, respectively, from above). For showing the approximation property, we assume to have another Hilbert space $X_{+,k}$ (playing the role of $H^2(\Omega)$ from above) such that the following conditions are satisfied:

1. The assumptions of the Babuška-Aziz theorem (Theorem 5): We need condition **(A1)**, introduced on page 17. We have to guarantee that this condition is also satisfied for the discretized problem. Therefore, we additionally need condition **(A1a)**, introduced on page 28.

2. We need the approximation error estimate **(A3)**, introduced below.

3. We need the regularity assumption **(A4)**, introduced below.

Here, for sake of self-containedness we give a complete proof of the approximation property using these assumptions. As we are interested in parameter-dependent problems, we need qualitative knowledge about the constants. Therefore, we show that only the constants in the conditions of the theorem affect the constant C_A in the approximation property.

Theorem 26 *Assume that there are Hilbert spaces* $X_{-,k} = (X_-, \|\cdot\|_{X_{-,k}})$, $X = (X, \|\cdot\|_X)$ *and* $X_{+,k} = (X_+, \|\cdot\|_{X_{+,k}})$ *such that* $X_+ \subseteq X \subseteq X_-$ *and the following conditions hold.*

(A1) *There are constants* $\underline{C} > 0$ *and* $\overline{C}$ *such that the well-posedness result*

$$\underline{C}\|x\|_X \leq \sup_{\tilde{x}\in X\setminus\{0\}} \frac{\mathcal{B}(x,\tilde{x})}{\|\tilde{x}\|_X} \leq \overline{C}\|x\|_X$$

holds for all $x \in X$.

(A1a) *There are constants* $\underline{C}_D > 0$ *and* $\overline{C}_D$ *such that for all grid levels* k *the well-posedness result for the discretized problem,*

$$\underline{C}_D\|x_k\|_X \leq \sup_{\tilde{x}_k\in X_k\setminus\{0\}} \frac{\mathcal{B}(x_k,\tilde{x}_k)}{\|\tilde{x}_k\|_X} \leq \overline{C}_D\|x_k\|_X$$

holds for all $x_k \in X_k$.

(A3) *There is a constant* $C_I > 0$ *such that for all grid levels* k *and all* $x \in X_+$ *the approximation error result*

$$\inf_{x_k\in X_k} \|x - x_k\|_X \leq C_I\|x\|_{X_{+,k}}$$

is satisfied.

(A4) *There is a constant* $C_R > 0$ *such that for all grid levels* k, *all* $\mathcal{F} \in (X_-)^*$ *the solution of the problem, find* $x_\mathcal{F} \in X$ *such that*

$$\mathcal{B}(x_\mathcal{F}, \tilde{x}) = \mathcal{F}(\tilde{x}) \qquad \text{for all } \tilde{x} \in X,$$

satisfies $x_\mathcal{F} \in X_+$ *and the bound*

$$\|x_\mathcal{F}\|_{X_{+,k}} \leq C_R\|\mathcal{F}\|_{(X_{-,k})^*}.$$

Then the approximation property (4.2) *is satisfied with a constant C_A, only depending on $\underline{C}$, $\overline{C}$, $\underline{C}_D$, $\overline{C}_D$, C_I and C_R*

Proof: The details of this proof follow Theorems 2.5 and 3.1 in Schöberl, Simon and Zulehner [53].

In this proof, for sake of simplicity C is a generic constant that only depends on $\underline{C}$, $\overline{C}$, $\underline{C}_D$, $\overline{C}_D$, C_I and C_R.

Let $x \in X$ and $x_k \in X_k$ be such that

$$\begin{aligned} \mathcal{B}(x, \tilde{x}) &= \mathcal{F}(\tilde{x}) && \text{for all } \tilde{x} \in X, \\ \mathcal{B}(x_k, \tilde{x}_k) &= \mathcal{F}(\tilde{x}_k) && \text{for all } \tilde{x}_k \in X_k. \end{aligned}$$

First we show that

$$\|x - x_k\|_{X_{-,k}} \leq C \sup_{\tilde{x} \in X_- \setminus \{0\}} \frac{\mathcal{F}(\tilde{x})}{\|\tilde{x}\|_{X_{-,k}}} \tag{4.10}$$

holds. The proof of this estimate follows the classical line of arguments: Because of **(A1)** and **(A1a)**, we can estimate the discretization error in the X-norm by the approximation error:

$$\|x - x_k\|_X \leq C \inf_{\tilde{x}_k \in X_k} \|x - \tilde{x}_k\|_X.$$

Using **(A3)** and **(A4)** we obtain further

$$\|x - x_k\|_X \leq C \|\mathcal{F}\|_{(X_{-,k})^*}.$$

For the estimate in the norm $\|\cdot\|_{X_{-,k}}$, we use the Aubin-Nitsche duality trick: For every (arbitrarily but fixed) $\mathcal{F}^* \in (X_-)^*$, we consider the following problem: Find $\hat{x}_{\mathcal{F}^*} \in X$ such that

$$\mathcal{B}(\tilde{x}, \hat{x}_{\mathcal{F}^*}) = \mathcal{F}^*(\tilde{x}) \qquad \text{for all } \tilde{x} \in X.$$

Using Galerkin orthogonality, we obtain

$$\mathcal{F}^*(x - x_k) = \mathcal{B}(x - x_k, \hat{x}_{\mathcal{F}^*}) = \mathcal{B}(x - x_k, \hat{x}_{\mathcal{F}^*} - \hat{x}_k)$$

for all $\hat{x}_k \in X_k$. Using **(A1)** and **(A1a)**, we obtain

$$\mathcal{F}^*(x - x_k) \leq C \|x - x_k\|_X \inf_{\hat{x}_k \in X_k} \|\hat{x}_{\mathcal{F}^*} - \hat{x}_k\|_X.$$

As above we obtain

$$\mathcal{F}^*(x - x_k) \leq C \|x - x_k\|_X \|\mathcal{F}^*\|_{(X_{-,k})^*},$$

which implies (as we may choose $\mathcal{F}^*$ arbitrarily)

$$\|x - x_k\|_{(X_{-,k})^*} = \sup_{\mathcal{F}^* \in (X_{-,k})^* \setminus \{0\}} \frac{\mathcal{F}^*(x - x_k)}{\|\mathcal{F}\|_{(X_{-,k})^*}} \leq C \|x - x_k\|_X,$$

which shows (4.10). Now we show the approximation property

$$\left\| x_k^{(1)} - x_k \right\|_{X_{-,k}} \leq C_A \sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\mathcal{B}\left(x_k^{(0,m)} - x_k, \tilde{x}_k\right)}{\|\tilde{x}_k\|_{X_{-,k}}}.$$

One easily sees that

$$x_k - x_k^{(1)} = t_k - t_{k-1},$$

with $t_k = x_k - x_k^{(0,m)}$ and $t_{k-1} \in X_{k-1}$ given by the formula for the coarse-grid correction step

$$\mathcal{B}(t_{k-1}, \tilde{x}_{k-1}) = \mathcal{F}(\tilde{x}_{k-1}) - \mathcal{B}\left(x_k^{(0,m)}, \tilde{x}_{k-1}\right) \qquad \text{for all } \tilde{x}_{k-1} \in X_{k-1}.$$

We observe that

$$\begin{aligned} \mathcal{B}(t_{k-1}, \tilde{x}_{k-1}) &= \mathcal{F}(\tilde{x}_{k-1}) - \mathcal{B}\left(x_k^{(0,m)}, \tilde{x}_{k-1}\right) \\ &= \mathcal{B}\left(x_k - x_k^{(0,m)}, \tilde{x}_{k-1}\right) \\ &= \mathcal{B}(t_k, \tilde{x}_{k-1}) \end{aligned} \tag{4.11}$$

for all $\tilde{x}_{k-1} \in X_{k-1}$. For a given $\mathcal{F}^* \in (X_{-,k})^*$, let $\hat{x} \in X$, $\hat{x}_k \in X_k$ and $\hat{x}_{k-1} \in X_{k-1}$ satisfy

$$\begin{aligned} \mathcal{B}(\tilde{x}, \hat{x}) &= \mathcal{F}^*(\tilde{x}) && \text{for all } \tilde{x} \in X, \\ \mathcal{B}(\tilde{x}_k, \hat{x}_k) &= \mathcal{F}^*(\tilde{x}_k) && \text{for all } \tilde{x}_k \in X_k, \\ \mathcal{B}(\tilde{x}_{k-1}, \hat{x}_{k-1}) &= \mathcal{F}^*(\tilde{x}_{k-1}) && \text{for all } \tilde{x}_{k-1} \in X_{k-1}. \end{aligned}$$

Then

$$\mathcal{F}^*(t_k - t_{k-1}) = \mathcal{B}(t_k - t_{k-1}, \hat{x}_k) = \mathcal{B}(t_k, \hat{x}_k - \hat{x}_{k-1})$$

since

$$\mathcal{B}(t_{k-1}, \hat{x}_k) = \mathcal{F}^*(t_{k-1}) = \mathcal{B}(t_{k-1}, \hat{x}_{k-1}) = \mathcal{B}(t_k, \hat{x}_{k-1})$$

using (4.11). Hence

$$\mathcal{F}^*(t_k - t_{k-1}) \leq \sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\mathcal{B}(t_k, \tilde{x}_k)}{\|\tilde{x}_k\|_{X_{-,k}}} \|\mathcal{F}^*\|_{(X_{-,k})^*}.$$

Therefore,

$$\|t_k - t_{k-1}\|_{X_{-,k}} = \sup_{\mathcal{F}^* \in (X_{-,k})^* \setminus \{0\}} \frac{\mathcal{F}^*(t_k - t_{k-1})}{\|\mathcal{F}^*\|_{(X_{-,k})^*}} \leq C \sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\mathcal{B}(t_k, \tilde{x}_k)}{\|\tilde{x}_k\|_{X_{-,k}}},$$

which completes the proof. □

Remark 27 *Assume that X is a dense subset of X_- and the solution of the problem, find $x_{\mathcal{F}} \in X$ such that*

$$\mathcal{B}(x_{\mathcal{F}}, \tilde{x}) = \mathcal{F}(\tilde{x}) \qquad \text{for all } \tilde{x} \in X,$$

satisfies $x_{\mathcal{F}} \in X_+$.

Then condition **(A4)** *is equivalent to*

$$C_R^{-1} \|x\|_{X_+,k} \leq \sup_{\tilde{x} \in X \setminus \{0\}} \frac{\mathcal{B}(x, \tilde{x})}{\|\tilde{x}\|_{X_-,k}} \qquad \text{for all } x \in X_+. \tag{4.12}$$

4.1.3 A two-grid convergence result

We can directly combine the smoothing property and the approximation property and obtain the following result.

Corollary 28 *Provided that both, the* smoothing property (4.1) *and the* approximation property (4.2)*, are satisfied, then*

$$\left\| x_k^{(1)} - x_k \right\|_{X_-,k} \leq C_A \, \eta(\nu) \left\| x_k^{(0)} - x_k \right\|_{X_-,k}$$

and $\lim_{\nu \to \infty} \eta(\nu) = 0$ holds, where $x_k^{(0)}$ is the initial guess, x_k is the exact solution and $x_k^{(1)}$ is the output of the two-grid iteration. Therefore, for ν large enough the two-grid method is a contraction with a contraction number bounded away from 1.

It is easy to extend this result to the two-grid method also having post-smoothing steps. Provided that the smoother satisfies the power-boundedness condition (4.7), one can show that the method converges if sufficiently many pre-smoothing steps are applied (no matter how many post-smoothing steps are applied). By considering the transpose of the iteration matrix, one can show that the two-grid method converges also if sufficiently many post-smoothing steps are applied (no matter how many pre-smoothing steps are applied).

4.1.4 A W-cycle multigrid convergence result

A perturbation argument can be used to show that the two-grid-convergence result implies the convergence of the W-cycle multigrid convergence.

Theorem 29 *Assume that we have a sequence of grids for $k \in \mathbb{N}$ with nested subspaces $X_0 \subseteq X_1 \subseteq X_2 \subseteq \cdots \subseteq X$. Assume that the multigrid method fits into the framework introduced in Section 3.1 and that the* smoothing property (4.1)*, the* approximation property (4.2)*, the* power-boundedness condition (4.7) *and the following condition hold.*

(A5) *There are constants $\underline{C}_C > 0$ and $\overline{C}_C$ such that on all grid levels k the estimate*

$$\underline{C}_C \|x_{k-1}\|_{X_{-,k-1}} \leq \|x_{k-1}\|_{X_{-,k}} \leq \overline{C}_C \|x_{k-1}\|_{X_{-,k-1}}$$

holds for all $x_{k-1} \in X_{k-1}$.

Then, for all $\nu \geq \nu_0$, the estimate

$$\left\| x_k^{(1)} - x_k \right\|_{X_{-,k}} \leq 2\, C_A\, \eta(\nu) \left\| x_k^{(0)} - x_k \right\|_{X_{-,k}}$$

holds, where $x_k^{(0)}$ is the initial guess, x_k is the exact solution and $x_k^{(1)}$ is the output of the W-cycle multigrid iteration. The constant ν_0 only depends on the constants in the approximation property, equation (4.7) *and condition* **(A5)** *and of the function $\eta(\nu)$.*

Therefore, for ν sufficiently large, the W-cycle multigrid method is a contraction with a contraction number bounded away from 1*.*

For a proof, see, e.g., Hackbusch [35], Theorem 7.12.

Condition **(A5)** is a weak assumption which can be guaranteed for all model problems we are discussing in this thesis.

As mentioned in Chapter 3, the W-cycle multigrid method is one possible realization of the multigrid method. An alternative would be the V-cycle multigrid method. In the numerical experiments for the model problems, the V-cycle multigrid method has shown the same convergence behavior as the W-cycle multigrid method. Since the complexity of one V-cycle is smaller than the complexity of one W-cycle by a certain factor, the V-cycle is method is always faster.

Nonetheless, a (rigorous) convergence proof for the V-cycle multigrid method is not known for saddle point problems, as the ideas of Theorem 29 cannot be carried over to the V-cycle. For elliptic problems, a V-cycle convergence analysis is available, see, e.g., Braess and Hackbusch [15]. Their proof relies to the energy norm which does not exist for indefinite problems. Therefore, to the knowledge of the author, convergence proofs for the V-cycle multigrid method for indefinite problems are not available.

4.2 Application to the model problems: non-robust convergence results

In this section we give a convergence proof for both, the reduced KKT-system and the original (non-reduced) KKT-system. We will start in the first subsection with the reduced KKT-system.

In the second subsection we will treat the non-reduced KKT-system. Although the non-reduced KKT-system characterizes the solution of the same model problem as the reduced KKT-system, this analysis is also of interest because sometimes a reduction to a 2-by-2 system is not possible or not reasonable, e.g., if the linearization of a non-linear problem or a problem with additional inequality constraints is considered. Moreover, we will see that for the non-reduced KKT-system some difficulty arises because the control variable u lives in $L^2(\Omega)$, not in $H^1(\Omega)$. We will see that the framework presented in this thesis can also be applied in this case.

Note that in the present section, the robustness in the parameter α is not an issue, therefore all constants arising in the analysis may depend on the choice of the parameter α.

4.2.1 An analysis for the reduced KKT-system

In this section we apply the theory introduced in Section 4.1 to Model Problem 1. We assume that $\alpha > 0$ is fixed. As already mentioned, the solution of the model problem is characterized by the reduced KKT-system, which reads as follows. Let $y_D \in L^2(\Omega)$. Find $(y, p) \in X = Y \times P = H^1(\Omega) \times H^1(\Omega)$ such that

$$\begin{array}{lcccll} (y,\tilde{y})_{L^2(\Omega_1)} & + & (\tilde{y},p)_{H^1(\Omega)} & = & (y_D,\tilde{y})_{L^2(\Omega)} & \text{for all } \tilde{y} \in Y, \\ (y,\tilde{p})_{H^2(\Omega)} & - & \alpha^{-1}(p,\tilde{p})_{L^2(\Omega_2)} & = & 0 & \text{for all } \tilde{p} \in P. \end{array}$$

For showing convergence, we analyze the conditions **(A1)** – **(A5)**.

Conditions (A1) and (A1a)

We have shown in Chapter 2 (Theorem 9) that the conditions **(A1)** and **(A1a)** are satisfied for the norm

$$\|x\|_X := \left(\|y\|_Y^2 + \|p\|_P^2\right)^{1/2},$$

where

$$\|y\|_Y := \|y\|_{H^1(\Omega)} \qquad \text{and} \qquad \|p\|_P := \|p\|_{H^1(\Omega)}.$$

Condition (A2)

The next step is to choose the norm $\|\cdot\|_{X_{-,k}}$ such that both, condition **(A2)**, introduced on page 52, is satisfied and that the matrix $\mathcal{L}_k$, representing that norm, can be easily inverted, at least in an approximative way. The latter is satisfied for (scaled) L^2-norms because the mass matrix, representing the L^2-norm, is spectrally equivalent to its diagonal under weak assumptions on the grid which are satisfied within our framework.

Therefore we choose $X_- := Y_- \times P_- := L^2(\Omega) \times L^2(\Omega)$ and use the following ansatz for the norm on X_-:

$$\|x\|_{X_{-,k}} := \left(\|y\|^2_{Y_{-,k}} + \|p\|^2_{P_{-,k}}\right)^{1/2},$$

where

$$\|y\|_{Y_{-,k}} := \eta_k \|y\|_{L^2(\Omega)} \qquad \text{and} \qquad \|p\|_{P_{-,k}} := \rho_k \|p\|_{L^2(\Omega)}.$$

For this choice, condition **(A2)** reads as follows:

$$\|y_k\|^2_{H^1(\Omega)} + \|p_k\|^2_{H^1(\Omega)} \leq \eta_k^2 \|y_k\|^2_{L^2(\Omega)} + \rho_k^2 \|p_k\|^2_{L^2(\Omega)} \text{ for all } (y_k, p_k) \in Y_k \times P_k.$$

These conditions can be shown using standard inverse inequalities if $\eta_k \geq C h_k^{-1}$ and $\rho_k \geq C h_k^{-1}$ for some constant C. Therefore, we choose the norm as follows:

$$\|x\|_{X_{-,k}} := \left(\|y\|^2_{Y_{-,k}} + \|p\|^2_{P_{-,k}}\right)^{1/2},$$

where

$$\|y\|_{Y_{-,k}} := h_k^{-1} \|y\|_{L^2(\Omega)} \qquad \text{and} \qquad \|p\|_{P_{-,k}} := h_k^{-1} \|p\|_{L^2(\Omega)}.$$

Condition (A4)

We discuss the condition **(A4)**, introduced on page 56, before discussing condition **(A3)**, because condition **(A4)** allows to motivate the choice of the norm $\|\cdot\|_{X_{+,k}}$.

For the discussion of the condition **(A4)** we need the following lemma.

Lemma 30 *Assume that the regularity assumption* **(R)**, *introduced on page 27, holds and let f and $g \in L^2(\Omega)$. Then the solution of the problem, find $(y, p) \in X = Y \times P$ such that*

$$\begin{array}{rcrcll} (y, \tilde{y})_{L^2(\Omega_1)} & + & (\tilde{y}, p)_{H^1(\Omega)} & = & (f, \tilde{y})_{L^2(\Omega)} & \text{for all } \tilde{y} \in Y, \\ (y, \tilde{p})_{H^2(\Omega)} & - & \alpha^{-1}(p, \tilde{p})_{L^2(\Omega_2)} & = & (g, \tilde{p})_{L^2(\Omega)} & \text{for all } \tilde{p} \in P \end{array} \tag{4.13}$$

holds, satisfies the following regularity result: $(y,p) \in H^2(\Omega) \times H^2(\Omega)$ *and*

$$\|y\|^2_{H^2(\Omega)} + \|p\|^2_{H^2(\Omega)} \leq C \left(\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Omega)} \right).$$

The constant C *only depends on* α *and the constants in assumptions* **(R)** *and* **(A1)**.

Proof: The first line of the KKT-system (4.13) can be rewritten as follows

$$(p,\tilde{y})_{H^1(\Omega)} = -(y,\tilde{y})_{L^2(\Omega_1)} + (f,\tilde{y})_{L^2(\Omega_1)} \qquad \text{for all } \tilde{y} \in Y = H^1(\Omega).$$

Assumption **(R)** states that $p \in H^2(\Omega)$ and

$$\|p\|_{H^2(\Omega)} \leq C_R \|f - y\|_{L^2(\Omega)} \leq C_R \left(\|f\|_{L^2(\Omega)} + \|y\|_Y \right).$$

Using the second line of the KKT-system, we obtain

$$\|y\|_{H^2(\Omega)} \leq C_R \left\| g - \alpha^{-1} p \right\|_{L^2(\Omega_2)} \leq C_R \max\{1, \alpha^{-1}\} \left(\|g\|_{L^2(\Omega)} + \|p\|_P \right).$$

Using condition **(A1)** and the fact that $\|\cdot\|_X \geq \|\cdot\|_{L^2(\Omega)}$, we obtain

$$\|x\|_X \leq \frac{1}{\underline{C}} \|\mathcal{F}\|_{X^*} \leq \frac{1}{\underline{C}} \|\mathcal{F}\|_{L^2(\Omega)}$$

for $\mathcal{F}(\cdot) := (f,\cdot)_{L^2(\Omega)} + (g,\cdot)_{L^2(\Omega)}$. This allows to conclude

$$\|y\|^2_{H^2(\Omega)} + \|p\|^2_{H^2(\Omega)} \leq C_R \max\{1, \alpha^{-1}\} \left(1 + \underline{C}^{-1}\right) \left(\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Omega)} \right),$$

which finishes the proof. □

For the model problem, condition **(A4)** reads as follows. There is a constant $C_R > 0$ such that for all grid levels k, all $f, g \in L^2(\Omega)$ the solution of the problem, find $x_{\mathcal{F}} \in X$ such that

$$\mathcal{B}(x_{\mathcal{F}}, \tilde{x}) = (f,\tilde{y})_{L^2(\Omega)} + (g,\tilde{y})_{L^2(\Omega)} \qquad \text{for all } \tilde{x} \in X,$$

satisfies $x_{\mathcal{F}} \in X_+$ and

$$\|x_{\mathcal{F}}\|_{X_{+,k}} \leq C_R \, h_k \left(\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Omega)} \right)^{1/2}.$$

Lemma 30 implies

$$h_k \left(\|y\|^2_{H^2(\Omega)} + \|p\|^2_{H^2(\Omega)} \right)^{1/2} \leq C \, h_k \left(\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Omega)} \right)^{1/2}.$$

This means that condition **(A4)** is satisfied for $X_+ := Y_+ \times P_+$ with norms

$$\|x\|_{X_{+,k}} := \left(\|y\|^2_{Y_{+,k}} + \|p\|^2_{P_{+,k}} \right)^{1/2},$$

where

$$\|y\|_{Y_{+,k}} := h_k \|y\|_{H^2(\Omega)} \qquad \text{and} \qquad \|p\|_{P_{+,k}} := h_k \|p\|_{H^2(\Omega)}.$$

Condition (A3)

For the model problem, condition **(A3)**, introduced on page 56, reads as follows. There is a constant $C_I > 0$ such that for all grid levels k and all $(y,p) \in Y_+ \times P_+ = H^2(\Omega) \times H^2(\Omega)$ the approximation error result

$$\inf_{(y_k,p_k)\in Y_k \times P_k} \left(\|y - y_k\|^2_{H^1(\Omega)} + \|p - p_k\|^2_{H^1(\Omega)} \right) \leq C_I^2 \, h_k^2 \, \left(\|y\|^2_{H^2(\Omega)} + \|p\|^2_{H^2(\Omega)} \right)$$

is satisfied.

This is a standard approximation error result, which immediately follows from the interpolation error estimate in Theorem 16.

Condition (A5)

As we are also interested in showing convergence of the W-cycle multigrid method, we have also to analyze condition **(A5)**, introduced on page 60, which reads as follows. There are constants $\underline{C}_C$ and $\overline{C}_C$ such that

$$\underline{C}_C \|x_{k-1}\|_{X_{-,k-1}} \leq \|x_{k-1}\|_{X_{-,k}} \leq \overline{C}_C \|x_{k-1}\|_{X_{-,k-1}} \qquad \text{for all } x_{k-1} \in X_{k-1}.$$

Here, the estimates for y_{k-1} and p_{k-1} easily decouple. Therefore we need

$$\underline{C}_C \|y_{k-1}\|_{Y_{-,k-1}} \leq \|y_{k-1}\|_{Y_{-,k}} \leq \overline{C}_C \|y_{k-1}\|_{Y_{-,k-1}} \qquad \text{for all } y_{k-1} \in Y_{k-1}$$

and the same result for the adjoined state p_{k-1}. Using the definition of the norms we can rewrite the condition from above

$$\underline{C}_C h_{k-1}^{-1} \|y_{k-1}\|_{L^2(\Omega)} \leq h_k^{-1} \|y_{k-1}\|_{L^2(\Omega)} \leq \overline{C}_C h_{k-1}^{-1} \|y_{k-1}\|_{L^2(\Omega)} \qquad \text{for all } y_{k-1} \in Y_{k-1},$$

which reduces to

$$\underline{C}_C \leq \frac{h_{k-1}}{h_k} \leq \overline{C}_C.$$

In case of uniform refinement, this is the case with constants $\underline{C}_C = \overline{C}_C = \frac{1}{2}$.

Convergence result

As we have shown the conditions **(A1)**, **(A1a)**, **(A3)** and **(A4)**, we can apply Theorem 26 and obtain the following result.

Corollary 31 *Consider Model Problem 1 and assume that the regularity assumption* **(R)** *is satisfied. Then the approximation property* (4.2) *holds with a constant C_A independent of the grid level. The constant C_A may depend on the choice of the parameter α.*

If the result is combined with a statement on the smoothing property, the convergence of the two-grid method follows. For the preconditioned normal equation smoother, we can combine the approximation property with Theorems 23 and 29 and obtain the following statement.

Corollary 32 *Consider Model Problem 1 and assume that the regularity assumption* **(R)** *is satisfied. Assume that the normal equation smoother (Subsections 3.2.1 and 4.1.4) is applied and that $\mathcal{L}_k$ and τ are chosen as mentioned in Corollary 25.*

Then there is a constant $C > 0$ independent of the grid level k such that

$$\left\| x_k^{(1)} - x_k \right\|_{X_{-,k}} \leq \frac{C}{\sqrt{\nu}} \left\| x_k^{(0)} - x_k \right\|_{X_{-,k}}$$

holds, where x_k is the exact solution, $x_k^{(0)}$ is the starting value and $x_k^{(1)}$ is the iterate after one step of the two-grid method or the W-cycle multigrid method.

Therefore, for ν large enough, the convergence rate is bounded away from 1 *by a constant independent of the grid level k. The convergence rate may depend on the choice of the parameter α.*

If the Courant element is chosen for discretization, an efficient implementation of the normal equation smoother for Model Problem 1 is possible using

$$\hat{\mathcal{L}}_k := \begin{pmatrix} \operatorname{diag} K_k & \\ & \operatorname{diag} K_k \end{pmatrix}. \tag{4.14}$$

For this choice of $\hat{\mathcal{L}}_k$, a refined analysis allows to compute how the damping parameter $\hat{\tau}$ has to be chosen such that the conditions of Corollary 25 and, as a consequence the conditions of Corollary 32, are satisfied.

Corollary 33 *For the choice* (4.14) *and*

$$\hat{\tau} \in \left(0\,,\, \frac{1}{2\left(1 + \max\left\{1, \alpha^{-1}\right\}\right)^2} \right), \tag{4.15}$$

the conditions of Corollary 25 are satisfied.

Proof: We show that the conditions of Corollary 25 are satisfied. The matrix $\mathcal{L}_k$ is given by

$$\mathcal{L}_k = \begin{pmatrix} h_k^{-2} M_k & \\ & h_k^{-2} M_k \end{pmatrix}.$$

Due to standard scaling arguments, this matrix is spectrally equivalent to $\hat{\mathcal{L}}_k$.

Because of the fact that the Courant element is chosen for discretization, the matrix K_k is diagonal dominant and therefore $2 \operatorname{diag} K_k \geq K_k \geq M_k$ holds. Therefore, the spectral radius of $\hat{\mathcal{L}}_k^{-1} \mathcal{A}_k$ can be estimated from above as follows.

$$\begin{aligned} \rho(\hat{\mathcal{L}}_k^{-1} \mathcal{A}_k) &= \rho \begin{pmatrix} \hat{K}_k^{-1} M_k \hat{K}_k^{-1/2} & \hat{K}_k^{-1/2} K_k \hat{K}_k^{-1/2} \\ \hat{K}_k^{-1} K_k \hat{K}_k^{-1/2} & -\alpha^{-1} \hat{K}_k^{-1/2} M_k \hat{K}_k^{-1/2} \end{pmatrix} \\ &\leq \rho \left(\hat{K}_k^{-1/2} K_k \hat{K}_k^{-1/2} \right) + \max \left\{ 1, \alpha^{-1} \right\} \rho \left(\hat{K}_k^{-1/2} M_k \hat{K}_k^{-1/2} \right) \\ &\leq 2 + 2 \max \left\{ 1, \alpha^{-1} \right\}, \end{aligned}$$

where $\hat{K}_k := \operatorname{diag} K_k$. This shows that (4.15) is an appropriate choice. □

The analysis presented in this section can be carried over to the *boundary control Model Problem 3* using the same norms as introduced above for Model Problem 1.

For showing the approximation property, we need a regularity assumption that includes boundary conditions. Therefore, we introduce the following assumption.

($\mathbf{R}_\Gamma$) There is a constant $C_R > 0$ such that the following result holds. For $f \in L^2(\Omega)$ and $g \in H^{1/2}(\partial\Omega)$ let $y_f \in H^1(\Omega)$ be the solution of

$$(y_f, \tilde{y})_{H^1(\Omega)} = (f, \tilde{y})_{L^2(\Omega)} + (g, \tilde{y})_{L^2(\partial\Omega)} \text{ for all } \tilde{y} \in H^1(\Omega).$$

Then $y_f \in H^2(\Omega)$ and

$$\|y_f\|_{H^2(\Omega)} \leq C_R \left(\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\Omega)} \right).$$

Based on this assumption we show the approximation property and – as a consequence – the convergence of the two-grid method and the W-cycle multigrid method.

Theorem 34 *Consider Model Problem 3 and assume that the domain has a sufficiently smooth boundary and that regularity assumption* **($\mathbf{R}_\Gamma$)** *is satisfied. Assume that the normal equation smoother (Subsections 3.2.1 and 4.1.4) is applied and that $\mathcal{L}_k$ and τ are chosen as mentioned in Corollary 25.*

Then there is a constant $C > 0$ independent of the grid level k such that

$$\left\| x_k^{(1)} - x_k \right\|_{X_{-,k}} \leq \frac{C}{\sqrt{\nu}} \left\| x_k^{(0)} - x_k \right\|_{X_{-,k}}$$

holds, where x_k is the exact solution, $x_k^{(0)}$ is the starting value and $x_k^{(1)}$ is the iterate after one step of the two-grid method or the W-cycle multigrid method.

Therefore, for ν large enough, the convergence rate is bounded away from 1 by a constant independent of the grid level k. The convergence rate may depend on the choice of the parameter α.

Proof: We have seen in Remark 10 that the conditions **(A1)** and **(A1a)** hold for Model Problem 3. Because the conditions **(A2)**, **(A3)** and **(A5)** do not depend on the bilinear form, they are also satisfied for Model Problem 3.

Therefore, only condition **(A4)** has to be shown.

Due to Adams and Fournier [1], Lemmas 7.40, 7.41 and Remark 7.45.1 on sufficiently smooth domains in $\mathbb{R}^2$, for $m > 1/2$ there is a trace operator T mapping $H^m(\Omega)$ to $H^{m-1/2}(\partial\Omega)$ such that

$$\|Ty\|_{H^{m-1/2}(\partial\Omega)} \leq C\|y\|_{H^m(\Omega)} \qquad \text{for all } y \in H^m(\Omega).$$

Using these results, we can show condition **(A4)** analogously to the proof of Lemma 30 as follows. When considering the problem

$$(y, \tilde{p})_{H^1(\Omega)} = \alpha^{-1}(p, \tilde{p})_{L^2(\partial\Omega)} + (g, \tilde{p})_{L^2(\Omega)} \qquad \text{for all } \tilde{p} \in P,$$

we first use the fact that the trace of p is in $H^{1/2}(\Omega)$ and use regularity assumption $(\mathbf{R}_\Gamma)$ afterwards to obtain

$$\begin{aligned} \|y\|_{H^2(\Omega)} &\leq C \max\left\{1, \alpha^{-1}\right\} \left(\|p\|_{H^{1/2}(\partial\Omega)} + \|g\|_{L^2(\Omega)} \right) \\ &\leq C \max\left\{1, \alpha^{-1}\right\} \left(\|p\|_{H^1(\Omega)} + \|g\|_{L^2(\Omega)} \right). \end{aligned}$$

The rest of the proof of condition **(A4)** is completely analogous to the proof of Lemma 30. This concludes the proof because the statement of the theorem follows from the conditions **(A1)** – **(A5)** using the Theorems 23, 26 and 29. □

The statement of Corollary 33 on an efficient implementation of the multigrid iteration for the Model Problem 1 is also satisfied for the boundary control Model Problem 3.

Note that here the analysis for Model Problem 3 was done for domains with a smooth boundary only. The generalization of such an analysis to convex polygonal domains is non-trivial. Due to GRISVARD [33], also for convex polygonal domains a H^2-regularity result can be shown, cf. Theorems 1.4.1 and 1.4.2 there. Fur this purpose, the traces on the line segments of the polygonal boundary have to be estimated individually. We do not comment on it in detail because the analysis is quite technically.

4.2.2 An analysis for the non-reduced KKT-system

In this subsection we discuss how to apply the convergence framework introduced in this thesis to the non-reduced KKT-system.

A convergence analysis in this case was already worked out in SIMON AND ZULEHNER [58] for Model Problem 2. In TAKACS AND ZULEHNER [61] it was shown that the analysis can be carried over to the boundary control Model Problem 3. In both cases the approximation property was shown using the framework introduced by BRENNER [22]. Here, we use the framework introduced in Section 4.1.

Here we work out the details for Model Problem 3. The details for Model Problem 1 can be worked out analogously. As already mentioned, the solution of the model problem is characterized by the (non-reduced) KKT-system, which reads as follows. Let $y_D \in L^2(\Omega)$. Find $(y,u,p) \in X = H^1(\Omega) \times L^2(\partial\Omega) \times H^1(\Omega)$ such that

$$\begin{array}{rcrcrcl} (y,\tilde{y})_{L^2(\Omega)} & & & + & (p,\tilde{y})_{H^1(\Omega)} & = & (y_D,\tilde{y})_{L^2(\Omega)} \\ & & \alpha(u,\tilde{u})_{L^2(\partial\Omega)} & - & (p,\tilde{u})_{L^2(\partial\Omega)} & = & 0 \\ (y,\tilde{p})_{H^1(\Omega)} & - & (u,\tilde{p})_{L^2(\partial\Omega)} & & & = & 0 \end{array}$$

for all $(\tilde{y},\tilde{u},\tilde{p}) \in X$. Again, we show that the conditions **(A1)** – **(A5)** hold.

Conditions (A1) and (A1a)

In Remark 11 we have seen that the conditions **(A1)** and **(A1a)** are satisfied for $(y,u) \in Y := H^1(\Omega) \times L^2(\partial\Omega)$ and $p \in P := H^1(\Omega)$ with norms

$$\|x\|_X := \left(\|(y,u)\|_Y^2 + \|p\|_P^2\right)^{1/2},$$

where

$$\|(y,u)\|_Y := \left(\|y\|_{H^1(\Omega)}^2 + \|u\|_{L^2(\partial\Omega)}^2\right)^{1/2} \qquad \text{and} \qquad \|p\|_P := \|p\|_{H^1(\Omega)}.$$

Condition (A2)

Analogously to the last subsection, a standard inverse inequality implies that condition **(A2)**, introduced on page 23, is satisfied for $(y,u) \in Y_- := L^2(\Omega) \times L^2(\partial\Omega)$ and $p \in P_- := L^2(\Omega)$ with norms

$$\|x\|_{X_{-,k}} := \Big(\|(y,u)\|^2_{Y_{-,k}} + \|p\|^2_{P_{-,k}} \Big)^{1/2},$$

where

$$\|(y,u)\|_{Y_{-,k}} := \Big(h_k^{-2}\|y\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\partial\Omega)} \Big)^{1/2} \qquad \text{and} \qquad \|p\|_{P_{-,k}} := h_k^{-1}\|p\|_{L^2(\Omega)}.$$

Condition (A3)

Analogously to the last subsection, we can show that

$$\inf_{x_k \in X_k} \|x - x_k\|_X \leq C_I \underbrace{\Big(h_k^2\|y\|^2_{H^2(\Omega)} + h_k^2\|u\|^2_{H^1(\partial\Omega)} + h_k^2\|p\|_{H^2(\Omega)} \Big)^{1/2}}_{\|(y,u,p)\|_{X_{\#,k}} :=} \tag{4.16}$$

holds. Certainly, also

$$\inf_{x_k \in X_k} \|x - x_k\|_X \leq \|x\|_X \tag{4.17}$$

holds because $0 \in X_k$. The combination of (4.17) and (4.17) implies

$$\inf_{x_k \in X_k} \|x - x_k\|_X \leq C_I \|x\|_{X + X_{\#,k}}. \tag{4.18}$$

So, for $X_{+,k} := X + X_{\#,k}$, the estimate (4.18) is exactly condition **(A3)**, as introduced on page 26.

Note that we can represent the Hilbert space $X_{+,k}$ explicitly: we have $X_+ = Y_+ \times P_+$, where $Y_+ = H^1(\Omega) \times L^2(\partial\Omega)$ and $P_+ = H^1(\Omega)$ with norms

$$\|x\|_{X_{+,k}} = \Big(\|(y,u)\|^2_{Y_{+,k}} + \|p\|^2_{P_{+,k}} \Big)^{1/2},$$

where

$$\|(y,u)\|_{Y_{+,k}} = \Big(\|y\|^2_{[h_k H^2(\Omega)] + H^1(\Omega)} + \|u\|^2_{[h_k H^1(\partial\Omega)] + L^2(\partial\Omega)} \Big)^{1/2}$$

and

$$\|p\|_{P_{+,k}} = \|p\|_{[h_k H^2(\Omega)] + H^1(\Omega)}.$$

Condition (A4)

The proof of condition **(A4)**, introduced on page 56, is a bit more involved than the proof for the case of the reduced KKT-system. Here, similar to the framework in BRENNER [22], we have to split the analysis into an analysis for the state variable y and the adjoined state p on the one hand and an analysis for the control u on the other hand and combine these results.

Lemma 35 *Assume that the regularity assumption* $\mathbf{(R_\Gamma)}$*, introduced on page 66, holds. Then the condition* **(A4)** *is satisfied in the framework of this subsection.*

Proof: In this proof C is a generic constant which is independent of h_k but may depend on α, C_R, $\underline{C}$ and $\overline{C}$.

As required, $\mathcal{F} \in (X_-)^* = L^2(\Omega) \times L^2(\partial\Omega) \times L^2(\Omega)$. Therefore, we can express $\mathcal{F}$ as follows:

$$\mathcal{F}(\tilde{x}) = \mathcal{F}_1(\tilde{x}) + \mathcal{F}_2(\tilde{x}),$$

where

$$\begin{aligned} \mathcal{F}_1(\tilde{y},\tilde{u},\tilde{p}) &= (f_1,\tilde{y})_{L^2(\Omega)} + (g,\tilde{p})_{L^2(\Omega)} && \text{for all } f_1 \in L^2(\Omega),\ g \in L^2(\Omega), \\ \mathcal{F}_2(\tilde{y},\tilde{u},\tilde{p}) &= (f_2,\tilde{u})_{L^2(\partial\Omega)} && \text{for all } f_2 \in L^2(\partial\Omega). \end{aligned}$$

Let $x_{\mathcal{F}_1}$ and $x_{\mathcal{F}_2}$ be such that

$$\begin{aligned} \mathcal{B}(x_{\mathcal{F}_1},\tilde{x}) &= \mathcal{F}_1(\tilde{x}) && \text{for all } \tilde{x} \in X, \\ \mathcal{B}(x_{\mathcal{F}_2},\tilde{x}) &= \mathcal{F}_2(\tilde{x}) && \text{for all } \tilde{x} \in X. \end{aligned}$$

Due to linearity, we have $x_{\mathcal{F}} = x_{\mathcal{F}_1} + x_{\mathcal{F}_2}$. So, it is sufficient to show

$$\|x_{\mathcal{F}_1}\|_{X_{\#,k}} \le \|\mathcal{F}_1\|_{X^*_{-,k}}, \tag{4.19}$$

$$\|x_{\mathcal{F}_2}\|_X \le \|\mathcal{F}_2\|_{X^*_{-,k}} \tag{4.20}$$

because

$$\begin{aligned} \|x_{\mathcal{F}}\|_{X_{\#,k}+X} &= \inf_{x_{\mathcal{F}}=x_1+x_2} \|x_1\|_{X_{\#,k}} + \|x_2\|_X \le \|x_{\mathcal{F}_1}\|_{X_{\#,k}} + \|x_{\mathcal{F}_2}\|_X \\ &\le C\left(\|\mathcal{F}_1\|_{X^*_{-,k}} + \|\mathcal{F}_2\|_{X^*_{-,k}}\right) \le C\left(\|\mathcal{F}_1\|^2_{X^*_{-,k}} + \|\mathcal{F}_2\|^2_{X^*_{-,k}}\right)^{1/2} = C\|\mathcal{F}\|_{X^*_{-,k}}, \end{aligned}$$

which is the statement we have to show.

First we show (4.19). If the right-hand side is chosen to be $\mathcal{F}_1$, the KKT-system can be rewritten as follows:

$$\begin{aligned} (p,\tilde{y})_{H^1(\Omega)} &= (f_1,\tilde{y})_{L^2(\Omega)} - (y,\tilde{y})_{L^2(\Omega)} && \text{for all } \tilde{y}\in H^1(\Omega)\\ \alpha^{-1}(u,\tilde{u})_{L^2(\partial\Omega)} &= (p,\tilde{u})_{L^2(\partial\Omega)} && \text{for all } \tilde{u}\in L^2(\Omega)\\ (y,\tilde{p})_{H^1(\Omega)} &= (g,\tilde{p})_{L^2(\Omega)} - (u,\tilde{p})_{L^2(\partial\Omega)} && \text{for all } \tilde{p}\in H^1(\Omega) \end{aligned} \tag{4.21}$$

The first line of the system (4.21) implies

$$\|p\|_{H^2(\Omega)} \leq C\left(\|f_1\|_{L^2(\Omega)} + \|y\|_{L^2(\Omega)}\right).$$

The second line of the system (4.21) implies $u = \alpha^{-1}Tp$, where T is the trace operator, and therefore

$$\|u\|_{H^1(\partial\Omega)} \leq C\alpha^{-1}\|p\|_{H^{3/2}(\Omega)} \leq C\alpha^{-1}\|p\|_{H^2(\Omega)}.$$

The third line of the system (4.21) implies

$$\|y\|_{H^2(\Omega)} \leq C\left(\|g\|_{L^2(\Omega)} + \|u\|_{H^{1/2}(\partial\Omega)}\right).$$

Using these results and the fact that $\|\mathcal{F}_1\|_{(X_{-,k})^*} = h_k\left(\|f_1\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Omega)}\right)^{1/2}$, we can show similar to the proof of Lemma 30 that (4.19) holds.

As $\|\mathcal{F}_2\|_{X^*_{-,k}} = \|f_2\|_{L^2(\partial\Omega)} = \|\mathcal{F}_2\|_{X^*}$ holds and condition **(A1)** is satisfied, also the estimate (4.20) holds, which finishes the proof. □

Condition (A5)

Condition **(A5)**, introduced on page 60, can be shown in the same way as in the last subsection.

Convergence result

As we have shown the conditions **(A1)** – **(A5)**, we can apply Theorems 23, 26 and 29 and obtain the following statement.

Corollary 36 *Consider Model Problem 3 and assume that the regularity assumption* $(\mathbf{R}_\Gamma)$ *is satisfied. Assume that the normal equation smoother (Subsections 3.2.1 and 4.1.4) is applied and that $\mathcal{L}_k$ and τ are chosen as mentioned in Corollary 25.*

Then there is a constant $C > 0$ independent of the grid level k such that

$$\left\|x_k^{(1)} - x_k\right\|_{X_{-,k}} \leq \frac{C}{\sqrt{\nu}}\left\|x_k^{(0)} - x_k\right\|_{X_{-,k}}$$

holds, where x_k is the exact solution, $x_k^{(0)}$ is the starting value and $x_k^{(1)}$ is the iterate after one step of the two-grid or the W-cycle multigrid method. The convergence rate may depend on the choice of the parameter α.

Therefore, for ν large enough, the convergence rate is bounded away from 1 by a constant independent of the grid level k.

If the Courant element is chosen for discretization, an efficient implementation of the normal equation smoother for Model Problem 3 is possible using

$$\hat{\mathcal{L}}_k := \begin{pmatrix} \operatorname{diag} K_k & & \\ & \operatorname{diag} M_{\Gamma,k} & \\ & & \operatorname{diag} K_k \end{pmatrix}. \tag{4.22}$$

If we fix this choice of $\hat{\mathcal{L}}_k$, a refined analysis allows to compute how the damping parameter $\hat{\tau}$ has to be chosen such that the conditions of Corollary 25 and, as a consequence the conditions of Corollary 36, are satisfied.

Corollary 37 *For the choice* (4.22) *and*

$$\hat{\tau} \in \left(0\,,\, \frac{1}{2(1+\max\{1,\alpha\})^2}\right),$$

the conditions of Corollary 25 are satisfied.

This corollary can be proven analogously to Corollary 33.

As mentioned above, an analysis for Model Problem 1 can be carried out in a completely analogous way. We skip that analysis because this does not give any deeper insight.

Remark 38 (Other smoothers) *Other smoothers, like Uzawa type smoothers discussed in* Simon and Zulehner *[58], or* Takacs and Zulehner *[61], also satisfy the smoothing property for both, the reduced KKT-system and the non-reduced KKT-system, for the model problems and the norms introduced in this section. Therefore also for these smoothers convergence can be guaranteed.*

4.3 Application to the model problem 2: a robust convergence result based on full regularity

In this section we are interested in convergence results which are robust in the parameter α, i.e., we want the convergence rate to be bounded away from 1 by a constant independent of the grid level k and the choice of the parameter α. Here, we have restrict ourselves to Model Problem 2. The statement we show in this section has already been carried out in Schöberl, Simon and Zulehner [53].

In this section we give the result in terms of the convergence framework introduced in Section 4.1. Here, we choose the norms slightly different to Schöberl, Simon and Zulehner [53] which will allow the extension of the convergence result to the partial elliptic regularity case, which will be carried out in Section 4.5. In this section, we follow the ideas introduced in Takacs and Zulehner [63]. For this purpose, we have to introduce the concept of interpolation spaces first.

4.3.1 Interpolation spaces

In this subsection we introduce interpolation spaces. We restrict ourselves to the definition and those results which are necessary for the convergence analysis for the model problem. For further information, we refer to standard text books, like Lions and Magenes [43], Butzer and Berens [26] and Adams and Fournier [1].

Let A_1 and A_2 be Hilbert spaces contained in a linear Hausdorff space. Between the Hilbert spaces $A_1 + A_2$ and $A_1 \cap A_2$, we introduce for every $\theta \in (0,1)$ interpolation spaces $[A_1, A_2]_\theta$ with norms

$$\|u\|^2_{[A_1,A_2]_\theta} = \int_0^\infty t^{-2\theta-1} \left(\inf_{u=u_1+u_2,\ u_1\in A_1,\ u_2\in A_2} \|u_1\|^2_{A_1} + t^2\|u_2\|^2_{A_2} \right) \mathrm{d}t. \qquad (4.23)$$

The interpolation space $[A_1, A_2]_\theta$ is the subset of A_1+A_2 of elements u with finite norm, i.e., where $\|u\|_{[A_1,A_2]_\theta} < \infty$. This definition refers to the K-method of constructing interpolation spaces, cf. the text books mentioned above. The space $[A_1, A_2]_\theta$ lies between $A_1 \cap A_2$ and $A_1 + A_2$, i.e., we have

$$A_1 \cap A_2 \subseteq [A_1, A_2]_\theta \subseteq A_1 + A_2$$

and

$$\|u\|_{A_1+A_2} \leq C(\theta)\|u\|_{[A_1,A_2]_\theta} \qquad \text{for all } u \in [A_1, A_2]_\theta$$

and

$$\|u\|_{[A_1,A_2]_\theta} \leq C(\theta)^{-1}\|u\|_{A_1\cap A_2} \qquad \text{for all } u \in A_1 \cap A_2, \tag{4.24}$$

where $C(\theta) = \sqrt{2\theta(1-\theta)}$, see (3.2.15) and (3.2.16) in Butzer and Berens [26]. Sobolev spaces with broken index, $H^\theta(\Omega)$ can be defined as the corresponding interpolation spaces, i.e., we have

$$H^{m+\theta}(\Omega) = [H^m(\Omega), H^{m+1}(\Omega)]_\theta$$

for all $\theta \in (0,1)$ and all $m \in \mathbb{Z}_0^+$. This definition is equivalent to other definitions of such Sobolev spaces, see, e.g., Theorem 4.3.6 in Butzer and Berens [26].

Such a statement is not only true for two consecutive standard Sobolev spaces $H^m(\Omega)$ and $H^{m+1}(\Omega)$, but also in general. This is a consequence of the following theorem.

Theorem 39 (Reiteration theorem) *Let A_1 and A_2 be Hilbert spaces. Then*

$$[[A_1, A_2]_{\theta_0}, [A_1, A_2]_{\theta_1}]_\lambda = [A_1, A_2]_{(1-\lambda)\theta_0+\lambda\theta_1}$$

and

$$[A_1, [A_1, A_2]_{\theta_1}]_\lambda = [A_1, A_2]_{\theta_1\lambda}$$

holds for all θ_0, θ_1, $\lambda \in (0,1)$.

For a proof, see Theorem 3.2.20 and Corollary 3.2.17 in Butzer and Berens [26].

Note that the second statement in Theorem 39 is not a consequence of the first statement in Theorem 39 because $[A_1, A_2]_{\theta_0}$ is not defined for $\theta_0 = 0$ in general.

Taking the dual space and interpolation between two Hilbert spaces commute, i.e., for Hilbert spaces A_1 and A_2, where $A_1 \cap A_2$ is dense in A_1 and in A_2, and all $\theta \in (0,1)$ the identity

$$[A_1^*, A_2^*]_\theta = ([A_1^*, A_2^*]_\theta)^*$$

holds (*duality theorem*), see Butzer and Berens [26], p. 214.

As we also work with weighted Sobolev spaces, which can be interpreted as intersections of the involved scaled spaces, we need some results on interpolation of intersections of Hilbert spaces and of scaled Hilbert spaces. Concerning scaling, interpolation spaces behave like the weighted geometric mean, i.e., for all Hilbert spaces A_1 and A_2 and all scalars α_1 and α_2 and all $\theta \in (0,1)$ the identity

$$[\alpha_1 A_1, \alpha_2 A_2]_\theta = \alpha_1^{1-\theta}\alpha_2^\theta [A_1, A_2]_\theta \tag{4.25}$$

holds. Due to monotonicity of the interpolation formula (4.23), we have for all Hilbert spaces A_1, A_2 and A_3, all $u \in [A_1 \cap A_2, A_3]_\theta$ and all $\theta \in (0,1)$.

$$\|u\|_{[A_1 \cap A_2, A_3]_\theta} \geq \|u\|_{[A_1, A_3]_\theta}$$

and therefore

$$\sqrt{2}\|u\|_{[A_1 \cap A_2, A_3]_\theta} \geq \|u\|_{[A_1, A_3]_\theta \cap [A_2, A_3]_\theta}. \tag{4.26}$$

A very powerful result for interpolation spaces, is the following statement.

Theorem 40 (Interpolation Theorem) *Let A_1, A_2, B_1 and B_2 be Hilbert spaces and let $T : A_1 + A_2 \to B_1 + B_2$ with $T(A_1) \subseteq B_1$ and $T(A_2) \subseteq B_2$ be an operator such that there are constants C_1 and C_2 such that*

$$\|Tu\|_{B_1} \leq C_1 \|x\|_{A_1} \text{ for all } u \in A_1,$$
$$\|Tu\|_{B_2} \leq C_2 \|x\|_{A_2} \text{ for all } u \in A_2.$$

Then the estimate

$$\|Tu\|_{[B_1, B_2]_\theta} \leq C(\theta)\, C_1^{1-\theta}\, C_2^{\theta} \|u\|_{[A_1, A_2]_\theta}$$

holds for all $u \in [A_1, A_2]_\theta$, where $C(\theta)$ only depends on θ.

For a proof, see Theorem 3.2.23 in Butzer and Berens [26].

4.3.2 An analysis for the reduced KKT-system

As already mentioned, the solution of the model problem is characterized by the reduced KKT-system, which reads as follows. Let $y_D \in L^2(\Omega)$. Find $(y, p) \in X = Y \times P = H^1(\Omega) \times H^1(\Omega)$ such that

$$\begin{array}{lcll} (y, \tilde{y})_{L^2(\Omega)} & + & (\tilde{y}, p)_{H^1(\Omega)} = (y_D, \tilde{y})_{L^2(\Omega)} & \text{for all } \tilde{y} \in Y, \\ (y, \tilde{p})_{H^2(\Omega)} & - & \alpha^{-1}(p, \tilde{p})_{L^2(\Omega)} = 0 & \text{for all } \tilde{p} \in P. \end{array}$$

Again, we discuss the conditions **(A1)** – **(A5)**.

Conditions (A1) and (A1a)

We have seen in Theorem 12 that the conditions **(A1)** and **(A1a)**, introduced on pages 17 and 28, are satisfied for $X = H^1(\Omega) \times H^1(\Omega)$ with norm

$$\|x\|_X = \left(\|y\|_Y^2 + \|p\|_P^2\right)^{1/2},$$

where

$$\|y\|_Y = \left(\|y\|^2_{L^2(\Omega)} + \alpha^{1/2}\|y\|^2_{H^1(\Omega)}\right)^{1/2}$$

and

$$\|p\|_P = \left(\alpha^{-1}\|p\|^2_{L^2(\Omega)} + \alpha^{-1/2}\|p\|^2_{H^1(\Omega)}\right)^{1/2}.$$

This result was already given in Schöberl and Zulehner [54]. In Zulehner [71] the same result was shown using condition **(A1')**, introduced on page 19.

Condition (A2)

Now we choose the norm $\|\cdot\|_{X_{-,k}}$ such that both, condition **(A2)**, introduced on page 52, is satisfied and that the matrix $\mathcal{L}_k$, representing that norm, can be easily inverted, at least in an approximative way. As mentioned in the last section, the latter is the case for (scaled) L^2-norms.

Therefore we use again the ansatz $X_- := Y_- \times P_- := L^2(\Omega) \times L^2(\Omega)$ and

$$\|x\|_{X_{-,k}} := \left(\|y\|^2_{Y_{-,k}} + \|p\|^2_{P_{-,k}}\right)^{1/2},$$

where

$$\|y\|_{Y_{-,k}} := \eta_k\|y\|_{L^2(\Omega)} \qquad \text{and} \qquad \|p\|_{P_{-,k}} := \rho_k\|p\|_{L^2(\Omega)}.$$

For this choice, condition **(A2)** reads as follows:

$$\|y_k\|^2_{L^2(\Omega)} + \alpha^{1/2}\|y_k\|^2_{H^1(\Omega)} + \alpha^{-1}\|p_k\|^2_{L^2(\Omega)} + \alpha^{-1/2}\|p_k\|^2_{H^1(\Omega)} \le \eta_k^2\|y_k\|^2_{L^2(\Omega)} + \rho_k^2\|p_k\|^2_{L^2(\Omega)}$$

for all $y_k \in Y_k$, $p_k \in P_k$. Due to inverse inequality it is sufficient to have $\eta_k^2 \ge C\,(1 + \alpha^{1/2}h_k^{-2})$ and $\rho_k^2 \ge C\,\alpha^{-1}\left(1 + \alpha^{1/2}h_k^{-2}\right)$. Therefore, we choose η_k and ρ_k as follows

$$\eta_k^2 := 1 + \alpha^{1/2}h_k^{-2} \qquad \text{and} \qquad \rho_k^2 := \alpha^{-1}\left(1 + \alpha^{1/2}h_k^{-2}\right). \tag{4.27}$$

Condition (A4)

We discuss condition **(A4)** before discussing condition **(A3)**, because condition **(A4)** allows to motivate the choice of the norm $\|\cdot\|_{X_{+,k}}$.

First note that we have seen in Subsection 4.2.1 (Lemma 30), that for $\mathcal{F} \in (X_-)^* = L^2(\Omega) \times L^2(\Omega)$, the corresponding solution $x_{\mathcal{F}} \in X_+ = H^2(\Omega) \times H^2(\Omega)$.

The estimate, we have shown in Subsection 4.2.1 was not robust in the parameter α. Therefore, we have to construct a robust estimate. We use Remark 27 and show that the inf-sup-condition (4.12) is satisfied.

For this purpose, we use the following theorem.

Theorem 41 *Assume that the following two conditions hold.*

(A4') *Assume that Y_+ is a dense subset of Y, Y is a dense subset of Y_-, P_+ is a dense subset of P and P is a dense subset of P_-. Assume that there is a value ψ_k (depending on k and α) such that*

$$0 \leq a(\cdot,\cdot)^{1/2} \leq C_{R3}\psi_k^{-1}\|\cdot\|_{Y_{-,k}} \leq C_{R4}\psi_k\|\cdot\|_{Y_{+,k}}$$

and

$$0 \leq c(\cdot,\cdot)^{1/2} \leq C_{R3}\psi_k^{-1}\|\cdot\|_{P_{-,k}} \leq C_{R4}\psi_k\|\cdot\|_{P_{+,k}}.$$

(A4") *Assume that there are constants $C_{R1} > 0$, $\overline{C}_{R1}$, $C_{R2} > 0$ and $\overline{C}_{R2}$ such that*

$$C_{R1}\|y\|_{Y_{+,k}} \leq \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{a(y,\tilde{y})}{\|\tilde{y}\|_{Y_{-,k}}} + \sup_{\tilde{p}\in P\setminus\{0\}} \frac{b(y,\tilde{p})}{\|\tilde{p}\|_{P_{-,k}}}$$

holds for all $y \in Y_+$ and

$$C_{R2}\|p\|^2_{P_{+,k}} \leq \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{b(\tilde{y},p)}{\|\tilde{y}\|_{Y_{-,k}}} + \sup_{\tilde{p}\in P\setminus\{0\}} \frac{c(p,\tilde{p})}{\|\tilde{p}\|_{P_{-,k}}}$$

holds for all $p \in P_+$.

Then condition (4.12) *holds and the constant in* (4.12) *only depends on the constants C_{R1} C_{R2} C_{R3} and C_{R4} in* **(A4')** *and* **(A4")**, *not on the choice of ψ_k.*

Proof: The proof is similar to the proofs of Theorem 2.3 in Zulehner [71].

We observe

$$\begin{aligned}\sup_{\tilde{x}\in X\setminus\{0\}}\frac{\mathcal{B}(x,\tilde{x})}{\|\tilde{x}\|_{X_{-,k}}} &\geq \frac{1}{2}\left(\sup_{\tilde{y}\in Y\setminus\{0\}}\frac{a(y,\tilde{y})+b(\tilde{y},p)}{\|\tilde{y}\|_{Y_{-,k}}}+\sup_{\tilde{p}\in P\setminus\{0\}}\frac{b(y,\tilde{p})-c(p,\tilde{p})}{\|\tilde{p}\|_{P_{-,k}}}\right)\\ &\geq \frac{1}{2}\left(\left(\sup_{\tilde{y}\in Y\setminus\{0\}}\frac{b(\tilde{y},p)}{\|\tilde{y}\|_{Y_{-,k}}}+\sup_{\tilde{p}\in P\setminus\{0\}}\frac{b(y,\tilde{p})}{\|\tilde{p}\|_{P_{-,k}}}\right)\right.\\ &\qquad\left.-\left(\sup_{\tilde{y}\in Y\setminus\{0\}}\frac{a(y,\tilde{y})}{\|\tilde{y}\|_{Y_{-,k}}}+\sup_{\tilde{p}\in P\setminus\{0\}}\frac{c(p,\tilde{p})}{\|\tilde{p}\|_{P_{-,k}}}\right)\right)\\ &= \frac{1}{2}(\xi-\eta)\|x\|_{X_{+,k}},\end{aligned}$$

where

$$\begin{aligned}\eta &:= \frac{\sup_{\tilde{y}\in Y\setminus\{0\}}\frac{a(y,\tilde{y})}{\|\tilde{y}\|_{Y_{-,k}}}+\sup_{\tilde{p}\in P\setminus\{0\}}\frac{c(p,\tilde{p})}{\|\tilde{p}\|_{P_{-,k}}}}{\|x\|_{X_{+,k}}}\\ \xi &:= \frac{\sup_{\tilde{y}\in Y\setminus\{0\}}\frac{b(\tilde{y},p)}{\|\tilde{y}\|_{Y_{-,k}}}+\sup_{\tilde{p}\in P\setminus\{0\}}\frac{b(y,\tilde{p})}{\|\tilde{p}\|_{P_{-,k}}}}{\|x\|_{X_{+,k}}}.\end{aligned}$$

A second bound follows from:

$$\begin{aligned}\sup_{\tilde{x}\in X\setminus\{0\}}\frac{\mathcal{B}(x,\tilde{x})}{\|\tilde{x}\|_{X_{-,k}}} &\geq \frac{\mathcal{B}((y,p),(y,-p))}{\|(y,-p)\|_{X_{-,k}}} = \frac{a(y,y)+c(p,p)}{\|(y,-p)\|_{X_{-,k}}} \geq \frac{a(y,y)+c(p,p)}{\psi^2\|(y,-p)\|_{X_{+,k}}}\\ &= \frac{\sup_{\tilde{y}\in Y\setminus\{0\}}\frac{a(y,\tilde{y})^2}{a(\tilde{y},\tilde{y})}+\sup_{\tilde{p}\in P\setminus\{0\}}\frac{c(p,\tilde{p})^2}{c(\tilde{p},\tilde{p})}}{\psi^2\|x\|_{X_{+,k}}}\\ &\geq \frac{\sup_{\tilde{y}\in Y\setminus\{0\}}\frac{a(y,\tilde{y})^2}{\|\tilde{y}\|^2_{Y_{-,k}}}+\sup_{\tilde{p}\in P\setminus\{0\}}\frac{c(p,\tilde{p})^2}{\|\tilde{p}\|^2_{P_{-,k}}}}{\|x\|_{X_{+,k}}} \geq \frac{1}{2}\eta^2\|x\|_{X_{+,k}}\end{aligned}$$

We know $\xi+\eta\geq\min\{C_{R1},C_{R2}\}>0$. In the same way as in Zulehner [71], we can show that there is a lower bound only depending on C_{R1}, C_{R2}, C_{R3} and C_{R4}. □

Remark 42 *One can show in the same way as it was done in* Zulehner *[71] that condition* (4.12) *implies condition* **(A4")**.

Still, our goal is to choose the norm $\|\cdot\|_{X_{+,k}}$ such that (4.12) (and therefore condition **(A4)**) is satisfied. Due to Theorem 41, it is sufficient to choose $\|\cdot\|_{X_{+,k}}$ such that the conditions **(A4')** and **(A4")** are satisfied. First we consider condition **(A4")**.

Using the definition of $\|\cdot\|_{X_{-,k}}$, the first line of condition **(A4")** reads as follows:

$$C_{R1}\|y\|_{Y_{+,k}} \leq \sup_{\tilde{y}\in Y\setminus\{0\}}\frac{(y,\tilde{y})_{L^2(\Omega)}}{\eta_k\|\tilde{y}\|_{L^2(\Omega)}}+\sup_{\tilde{p}\in P\setminus\{0\}}\frac{(y,\tilde{p})_{H^1(\Omega)}}{\rho_k\|\tilde{p}\|_{L^2(\Omega)}} \quad \text{for all } y\in Y_+, \tag{4.28}$$

where η_k and ρ_k are given by (4.27). Now, we use equation (4.28) to derive the norm $\|\cdot\|_{Y_{+,k}}$. It is easy to see, that

$$\sup_{\tilde{y}\in Y\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega)}}{\eta_k\|\tilde{y}\|_{L^2(\Omega)}} = \eta_k^{-1}\|y\|_{L^2(\Omega)}$$

is satisfied. Due to Remark 20, there are constants $\underline{C}_R > 0$ and $\overline{C}_R$ such that

$$\underline{C}_R\,\rho_k^{-1}\,\|y\|_{H^2(\Omega)} \le \sup_{\tilde{p}\in P\setminus\{0\}} \frac{(y,\tilde{p})_{H^1(\Omega)}}{\rho_k\|\tilde{p}\|_{L^2(\Omega)}} \le \overline{C}_R\,\rho_k^{-1}\,\|y\|_{H^2(\Omega)}$$

holds for all $y \in Y_+$. So, condition (4.28) simplifies to: Guarantee that there is a constant $C_{R1} > 0$ such that

$$C_{R1}\|y\|_{Y_{+,k}} \le \eta_k^{-1}\|y\|_{L^2(\Omega)} + \underline{C}_R\rho_k^{-1}\|y\|_{H^2(\Omega)} \qquad \text{for all } y \in Y_+,$$

where ρ_k and η_k are given by (4.27). We can do the same for the second line of condition **(A4")**: Guarantee that there is a constant $C_{R2} > 0$ such that

$$C_{R2}\|p\|_{P_{+,k}} \le \alpha^{-1}\rho_k^{-1}\|p\|_{L^2(\Omega)} + \underline{C}_R\eta_k^{-1}\|p\|_{H^2(\Omega)} \qquad \text{for all } p \in P_+.$$

This motivates to choose $X_+ := Y_+ \times P_+ := H^2(\Omega) \times H^2(\Omega)$ with associated norms

$$\|x\|_{X_{+,k}} := \left(\|y\|^2_{Y_{+,k}} + \|p\|^2_{P_{+,k}}\right)^{1/2},$$

where

$$\|y\|_{Y_{+,k}} := \left(1+\alpha^{1/2}h_k^{-2}\right)^{-1/2}\left(\|y\|^2_{L^2(\Omega)} + \alpha\|y\|^2_{H^2(\Omega)}\right)^{1/2}$$

and

$$\|p\|_{P_{+,k}} := \alpha^{-1}\left(1+\alpha^{1/2}h_k^{-2}\right)^{-1/2}\left(\|p\|^2_{L^2(\Omega)} + \alpha\|p\|^2_{H^2(\Omega)}\right)^{1/2}.$$

By construction, condition **(A4")** is satisfied with constants C_{R1} and C_{R2} only depending on $\underline{C}_R$. It is easy to see, that the condition **(A4')** is satisfied with $C_{R3} = C_{R4} = 1$ for $\psi_k := \left(1+\alpha^{1/2}h_k^{-2}\right)^{1/2}$ because $\|\cdot\|_{H^2(\Omega)} \ge \|\cdot\|_{L^2(\Omega)}$ holds. Therefore, Theorem 41 implies that condition **(A4)** holds.

We observe that, for all grid levels k, the Hilbert space $(X, \|\cdot\|_X)$ is the interpolant between $(X_-, \|\cdot\|_{X_{-,k}})$ and $(X_+, \|\cdot\|_{X_{+,k}})$ at $\frac{1}{2}$, i.e.,

$$X = [(X_{-,k}), (X_{+,k})]_{1/2}. \tag{4.29}$$

Note that this is also satisfied in classical proof for the Laplace equation, cf Hackbusch [35]. There the norms are – in the notation of the present thesis – $\|\cdot\|_{X_{-,k}} = h_k^{-1}\|\cdot\|_{L^2(\Omega)}$, $\|\cdot\|_X^s = \|\cdot\|_{H^1(\Omega)}$ and $\|\cdot\|_{X_{+,k}} = h_k\|\cdot\|_{H^2(\Omega)}$.

We want to mention that in SCHÖBERL, SIMON AND ZULEHNER [53] the norm $\|\cdot\|_{X_{+,k}}$ was defined in a different way, as it is done in this thesis. There, the relation (4.29) was not satisfied. In Section 4.5 we will see that this relation can be used to construct the spaces needed for the partial regularity case in a straight-forward way.

Condition (A3)

The following lemma shows condition **(A3)**, as introduced on page 56.

Lemma 43 *There is a constant $C_I > 0$ such that*

$$\inf_{(y_k,p_k)\in Y_k\times P_k} \left(\|y-y_k\|_Y^2 + \|p-p_k\|_P^2\right)^{1/2} \leq C_I \left(\|y\|_{Y_{+,k}} + \|p\|_{P_{+,k}}\right)^{1/2}$$

holds for all $y \in Y_+$ and $p \in P_+$.

Proof: Throughout this proof, C is a generic constant that only depends on the constant in Theorem 16.

Due to Theorem 16 there is an interpolation operator such that for all $y \in H^2(\Omega)$

$$\begin{aligned} \|y-\Pi_k y\|_{L^2(\Omega)}^2 &\leq C\|y\|_{L^2(\Omega)}^2 \qquad & \|y-\Pi_k y\|_{L^2(\Omega)}^2 &\leq Ch_k^2\|y\|_{H^1(\Omega)}^2 \\ \|y-\Pi_k y\|_{H^1(\Omega)}^2 &\leq C\|y\|_{H^1(\Omega)}^2 \qquad & \|y-\Pi_k y\|_{H^1(\Omega)}^2 &\leq Ch_k^2\|y\|_{H^2(\Omega)}^2 \end{aligned}$$

holds.

Moreover we find out that for all Hilbert spaces A_1 and A_2 the relation $\|\cdot\|_{[A_1,A_2]_{1/2}} \leq \sqrt{2}\|\cdot\|_{A_1\cap A_2}$ holds. Therefore, the following inequality holds:

$$\alpha^{1/2}\|y\|_{H^1(\Omega)}^2 \leq 2\left(\|y\|_{L^2(\Omega)}^2 + \alpha\|y\|_{H^2(\Omega)}^2\right).$$

So we have

$$\begin{aligned} &\left(1+\alpha^{1/2}h_k^{-2}\right)\|y-\Pi_k y\|_Y^2 \\ &\quad= \left(1+\alpha^{1/2}h_k^{-2}\right)\left(\|y-\Pi_k y\|_{L^2(\Omega)}^2 + \alpha^{1/2}\|y-\Pi_k y\|_{H^1(\Omega)}^2\right) \\ &\quad= \|y-\Pi_k y\|_{L^2(\Omega)}^2 + \alpha^{1/2}h_k^{-2}\|y-\Pi_k y\|_{L^2(\Omega)}^2 \\ &\qquad\quad + \alpha^{1/2}\|y-\Pi_k y\|_{H^1(\Omega)}^2 + \alpha h_k^{-2}\|y-\Pi_k y\|_{H^1(\Omega)}^2 \\ &\quad\leq C\left(\|y\|_{L^2(\Omega)}^2 + \alpha^{1/2}\|y\|_{H^1(\Omega)}^2 + \alpha^{1/2}\|y\|_{H^1(\Omega)}^2 + \alpha\|y\|_{H^2(\Omega)}^2\right) \\ &\quad\leq C\left(\|y\|_{L^2(\Omega)}^2 + \alpha\|y\|_{H^2(\Omega)}^2\right) \\ &\quad= C\left(1+\alpha^{1/2}h_k^{-2}\right)\|y\|_{Y_{+,k}}^2. \end{aligned}$$

As we can show the analogous result for $\|\cdot\|_P$ and $\|\cdot\|_{P_{+,k}}$, the desired result follows immediately. □

Condition (A5)

Similar to last section, condition **(A5)**, introduced on page 60, reduces for quasi-uniform grids to show

$$\underline{C}_C^2 \leq \frac{1+\alpha^{1/2}h_{k-1}^{-2}}{1+\alpha^{1/2}h_k^{-2}} \leq \overline{C}_C^2.$$

This holds in a standard setting for $\underline{C}_C = \frac{1}{2}$ and $\overline{C}_C = 1$.

Convergence result

As we have shown the conditions **(A1)**, **(A1a)**, **(A3)** and **(A4)**, we can apply Theorem 26 and obtain the following result.

Corollary 44 *Consider Model Problem 2 and assume that the regularity assumption* **(R)** *is satisfied. Then the approximation property holds with a constant C_A independent of the grid level k and the choice of the parameter α.*

If the result is combined with a statement on the smoothing property, the convergence of the two-grid method follows. For the preconditioned normal equation smoother, we can combine the approximation property with Theorems 23 and 29 and obtain the following statement.

Corollary 45 *Consider Model Problem 2, assume that the regularity assumption* **(R)** *is satisfied. Assume that the normal equation smoother (Subsections 3.2.1 and 4.1.4) is applied and that $\mathcal{L}_k$ and τ are chosen as mentioned in Corollary 25.*

Then there is a constant $C > 0$ independent of the grid level k and the choice of the parameter α such that

$$\left\| x_k^{(1)} - x_k \right\|_{X_{-,k}} \leq \frac{C}{\sqrt{\nu}} \left\| x_k^{(0)} - x_k \right\|_{X_{-,k}} \tag{4.30}$$

holds, where x_k is the exact solution, $x_k^{(0)}$ is the starting value and $x_k^{(1)}$ is the iterate after one step of the two-grid or the W-cycle multigrid method.

Therefore, for ν large enough, the convergence rate is bounded away from 1 by a constant independent of the grid level k and the parameter α.

The convergence result is a statement in non-standard mesh-dependent norm $\|\cdot\|_{X_{-,k}}$. The following corollary shows that this result also implies convergence (with the same rate) in the standard norm $\|\cdot\|_{L^2(\Omega)}$.

Corollary 46 *Under the notations and assumptions of Corollary 45 there is a constant $C > 0$ and a factor $q = \frac{C}{\sqrt{\nu}}$, both independent of k and α, such that the L^2-convergence result*

$$\left\| x_k^{(n)} - x_k \right\|_{L^2(\Omega)} \leq C q^n \|y_D\|_{L^2(\Omega)}$$

holds for all $n \in \mathbb{N}$ and all $\alpha \in (0, 1]$, provided $y_k^{(0)} = p_k^{(0)} = 0$.

Proof: This proof was published in Takacs and Zulehner [62].

Corollary 45 states (4.30), which is equivalent to

$$\left(\left\| y_k^{(n)} - y_k \right\|^2_{L^2(\Omega)} + \alpha^{-1} \left\| p_k^{(n)} - p_k \right\|^2_{L^2(\Omega)} \right)^{1/2} \leq q^n \left(\left\| y_k^{(0)} - y_k \right\|^2_{L^2(\Omega)} + \alpha^{-1} \left\| p_k^{(0)} - p_k \right\|^2_{L^2(\Omega)} \right)^{1/2}.$$

Assuming $y_k^{(0)} = p_k^{(0)} = 0$ implies

$$\left(\left\| y_k^{(n)} - y_k \right\|^2_{L^2(\Omega)} + \alpha^{-1} \left\| p_k^{(n)} - p_k \right\|^2_{L^2(\Omega)} \right)^{1/2} \leq q^n \left(\|y_k\|^2_{L^2(\Omega)} + \alpha^{-1} \|p_k\|^2_{L^2(\Omega)} \right)^{1/2},$$

The right-hand-side is bounded from above by $q^n \|x_k\|_X$. Using **(A1a)**, we obtain

$$\|x_k\|_X \leq \underline{C}^{-1} \sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\mathcal{B}(x_k, \tilde{x}_k)}{\|\tilde{x}_k\|_X} = \underline{C}^{-1} \sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\mathcal{F}(\tilde{x}_k)}{\|\tilde{x}_k\|_X}.$$

Using

$$\mathcal{F}(\tilde{y}_k, \tilde{p}_k) = (y_D, \tilde{y}_k)_{L^2(\Omega)} \leq \|y_D\|_{L^2(\Omega)} \|\tilde{y}_k\|_{L^2(\Omega)} \leq \|y_D\|_{L^2(\Omega)} \|(\tilde{y}_k, \tilde{p}_k)\|_X,$$

we obtain $\|x_k\|_X \leq \underline{C}^{-1} \|y_D\|_{L^2(\Omega)}$ and further

$$\left(\left\| y_k^{(n)} - y_k \right\|^2_{L^2(\Omega)} + \alpha^{-1} \left\| p_k^{(n)} - p_k \right\|^2_{L^2(\Omega)} \right)^{1/2} \leq \underline{C}^{-1} q^n \|y_D\|_{L^2(\Omega)}.$$

For $\alpha \leq 1$, we have

$$\left(\left\| y_k^{(n)} - y_k \right\|^2_{L^2(\Omega)} + \left\| p_k^{(n)} - p_k \right\|^2_{L^2(\Omega)} \right)^{1/2} \leq \left(\left\| y_k^{(n)} - y_k \right\|^2_{L^2(\Omega)} + \alpha^{-1} \left\| p_k^{(n)} - p_k \right\|^2_{L^2(\Omega)} \right)^{1/2},$$

which completes the proof. □

If the Courant element is chosen for discretization, an efficient implementation of the normal equation smoother for Model Problem 1 is possible using

$$\hat{\mathcal{L}}_k := \begin{pmatrix} \operatorname{diag}(M_k + \alpha^{1/2} K_k) & \\ & \alpha^{-1} \operatorname{diag}(M_k + \alpha^{1/2} K_k) \end{pmatrix}. \tag{4.31}$$

If we fix this choice of $\hat{\mathcal{L}}_k$, a refined analysis allows to compute how the damping parameter $\hat{\tau}$ has to be chosen such that the conditions of Corollary 25 and, as a consequence the conditions of Corollary 45, are satisfied.

Corollary 47 *For the choice* (4.31) *and*

$$\hat{\tau} \in \left(0, \frac{1}{8}\right). \tag{4.32}$$

the conditions of Corollary 25 are satisfied.

Proof: The matrix $\mathcal{L}_k$ is given by

$$\mathcal{L}_k = \begin{pmatrix} (1 + \alpha^{1/2} h_k^{-2}) M_k & \\ & \alpha^{-1}(1 + \alpha^{1/2} h_k^{-2}) M_k \end{pmatrix}.$$

Due to standard scaling arguments, this matrix is spectrally equivalent to $\hat{\mathcal{L}}_k$, specified in (4.31).

Let $\hat{A}_k := \operatorname{diag}\left(M_k + \alpha^{1/2} K_k\right)$. Since the Courant element is chosen for discretization, the matrix K_k is diagonal dominant and therefore $2 \operatorname{diag} K_k \geq K_k$ and $2 \operatorname{diag} M_k \geq M_k$ holds. Therefore, the spectral radius of $\hat{\mathcal{L}}_k^{-1} \mathcal{A}_k$ can be estimated from above as follows.

$$\begin{aligned} \rho(\hat{\mathcal{L}}_k^{-1} \mathcal{A}_k) &= \rho \begin{pmatrix} \hat{A}_k^{-1} M_k \hat{A}_k^{-1/2} & \alpha^{1/2} \hat{A}_k^{-1/2} K_k \hat{A}_k^{-1/2} \\ \alpha^{1/2} \hat{A}_k^{-1} K_k \hat{A}_k^{-1/2} & -\hat{A}_k^{-1/2} M_k \hat{A}_k^{-1/2} \end{pmatrix} \\ &\leq \alpha^{1/2} \rho\left(\hat{A}_k^{-1/2} K_k \hat{A}_k^{-1/2}\right) + \rho\left(\hat{A}_k^{-1/2} M_k \hat{A}_k^{-1/2}\right) \\ &\leq 4, \end{aligned}$$

which shows that for (4.32) the conditions of Corollary 25 are satisfied. □

Remark 48 (Other smoothers) *We want to mention, that other smoothers, like Uzawa type smoothers discussed in* Schöberl, Simon and Zulehner *[53] and collective point smoothers which will be discussed in the next section, also satisfy the smoothing property for the model problem and the norms introduced in this section. Therefore also for these smoothers convergence can be guaranteed.*

4.4 Smoothing property for collective point smoothers and its application to the model problem 2

In the last section, we gave a rigorous convergence result for the case that the normal equation smoother is chosen. In this section, we show that the convergence results also holds if a collective point smoother is used. Here, we have to restrict to the case of collective Richardson smoothers. In Theorem 49 we give a smoothing result that relies on algebraic relations between the involved matrices. In Corollary 50, we see that this theorem can be applied in the framework of this thesis for Model Problem 2.

As already mentioned, the collective point smoothers are very popular and have been proposed, e.g., in Trottenberg [66], Borzi, Kunisch and Kwak [12], Borzi and Schulz [13] and Lass [41]. Convergence analysis based on Fourier analysis was available, but a rigorous convergence analysis was, up to the author's knowledge, not available. The convergence theorem presented in this section was worked out an and published in Takacs and Zulehner [62].

Theorem 49 *Consider the block-matrix $\mathcal{A}_k$, which is given by*

$$\mathcal{A}_k = \begin{pmatrix} A_k & B_k \\ B_k & -\alpha^{-1} A_k \end{pmatrix},$$

where $A_k, B_k \in \mathbb{R}^{N_k \times N_k}$ are symmetric and positive definite matrices. Let the preconditioner $\hat{\mathcal{A}}_k$ be given by

$$\hat{\mathcal{A}}_k := \begin{pmatrix} \hat{A}_k & \hat{B}_k \\ \hat{B}_k & -\alpha^{-1} \hat{A}_k \end{pmatrix}.$$

Here, $\hat{A}_k, \hat{B}_k \in \mathbb{R}^{N_k \times N_k}$ are preconditioners such that

$$\rho\left(I - \hat{A}_k^{-1} A_k\right) \leq 1 \qquad \text{and} \qquad \rho\left(I - \hat{B}_k^{-1} B_k\right) \leq 1 \tag{4.33}$$

holds. Moreover we assume that there is a symmetric positive definite matrix $\hat{D}_k$ such that $\hat{A}_k := a_k \hat{D}_k$ and $\hat{B}_k := b_k \hat{D}_k$, where $a_k > 0$ and $b_k > 0$ are scalars.

Then, for all $\tau \in (0,1)$, there is a constant $C_S > 0$ such that

$$\left\| \mathcal{L}_k^{-1/2} \mathcal{A}_k \left(I - \tau \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k\right)^{\nu} \mathcal{L}_k^{-1/2} \right\|_{\ell^2} \leq \frac{C_S}{\sqrt{\nu}}$$

holds for all grid levels $k \in \{0, \dots, K\}$, for all choices of $\alpha > 0$ and for all $\nu \in \mathbb{N}$. Here, the matrix $\mathcal{L}_k$ is given by

$$\mathcal{L}_k := \begin{pmatrix} (\hat{A}_k^2 + \alpha \hat{B}_k^2)^{1/2} & \\ & \alpha^{-1}(\hat{A}_k^2 + \alpha \hat{B}_k^2)^{1/2} \end{pmatrix}.$$

Moreover, the iteration scheme is power-bounded, i.e.,

$$\left\| \mathcal{L}_k^{1/2} \left(I - \tau \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k \right)^{\nu} \mathcal{L}_k^{-1/2} \right\|_{\ell^2} \leq 2$$

holds for all $\nu \in \mathbb{N}$.

Before we prove this theorem, we discuss its application to the Model Problem 2.

Corollary 50 *Consider the reduced KKT-system for Model Problem 2. Then the collective Richardson iteration, introduced in Subsection 3.2.2 satisfies for $\tau \in (0, 1)$ the smoothing property with smoothing rate*

$$\eta(\nu) = \frac{C_S}{\sqrt{\nu}},$$

where the constant C_S is independent of the grid level k and the choice of the parameter α.

If moreover the regularity assumption **(R)**, *introduced on page 27, is satisfied, there is a constant $C > 0$ independent of the grid level k such that*

$$\left\| x_k^{(1)} - x_k \right\|_{X_{-,k}} \leq \frac{C}{\sqrt{\nu}} \left\| x_k^{(0)} - x_k \right\|_{X_{-,k}}$$

holds, where x_k is the exact solution, $x_k^{(0)}$ is the starting value and $x_k^{(1)}$ is the iterate after one step of the two-grid or the W-cycle multigrid method. The convergence rate may depend on the choice of the parameter α.

Therefore, for ν large enough, the convergence rate is bounded away from 1 by a constant independent of the grid level k.

Proof: Here, we use Theorem 49 with $\hat{D}_k = I$ and use the fact that $\mathcal{L}_k$, introduced in Theorem 49, and $\mathcal{L}_k$, representing the norm $|||\cdot|||_{0,k}$, introduced in Section 4.3, are spectrally equivalent.

This shows the smoothing property. The approximation property was shown in Section 4.3. Therefore, the convergence of the two-grid method and the W-cycle multigrid method immediately follow. □

For the proof of Theorem 49 we use a variant of Reusken's lemma. See REUSKEN [52] for the original work.

Lemma 51 *Let $\mathcal{L}_k$ be a symmetric positive definite matrix and let $\mathcal{M}_k$ be a matrix that is power bounded with respect to $\|\cdot\|_{\mathcal{L}_k}$, i.e., there is a constant C_B such that*

$$\|\mathcal{M}_k^\nu \underline{x}_k\|_{\mathcal{L}_k} \leq C_B \|\underline{x}\|_{\mathcal{L}_k} \tag{4.34}$$

for all $\nu \in \mathbb{N}$.

Then for every choice of the damping parameter $\tau \in (0,1)$ there is a constant C (independent of h_k and α) such that

$$\|(I - \mathcal{M}_k)((1-\tau)I + \tau\mathcal{M}_k)^\nu\|_{\mathcal{L}_k} \leq \frac{C}{\sqrt{\nu}}$$

holds for all $\nu \in \mathbb{N}$.

Proof: The proof was given in ECKER AND ZULEHNER [31] for the case $\|\mathcal{M}_k\|_{\mathcal{L}_k} \leq 1$ and can easily be extended to the case that $\mathcal{M}_k$ is power bounded. □

Remark 52 *Note that the fact that a linear iteration with iteration matrix $\mathcal{M}_k$ is power boundedness implies that also the damped iteration with damping parameter $\tau \in (0,1)$ and iteration matrix $(1-\tau)I + \tau\mathcal{M}_k$ is power bounded.*

Due to Lemma 51, we have to show that the iteration matrix of the (non-damped) iteration scheme is power bounded. This will be done in the next two lemmas.

Lemma 53 *Using the notations of Theorem 49, the identity*

$$\|(I - \hat{\mathcal{A}}_k^{-1}\mathcal{A}_k)^\nu\|_{\mathcal{L}_k} = \|\tilde{Z}_k^\nu\|_{\ell^2}$$

holds for all $\nu \in \mathbb{N}$, where $\tilde{Z}_k$ is given by

$$\tilde{Z}_k := (\hat{A}_k^2 + \alpha\hat{B}_k^2)^{1/4}\, Z_k\, (\hat{A}_k^2 + \alpha\hat{B}_k^2)^{-1/4}$$

with

$$Z_k := (\hat{A}_k + \sqrt{\alpha}\hat{B}_k\mathrm{i})^{-1}(\Delta A_k + \sqrt{\alpha}\Delta B_k\mathrm{i}).$$

Proof: One easily verifies that

$$I - \hat{\mathcal{A}}_k^{-1}\mathcal{A}_k = \hat{\mathcal{A}}_k^{-1}\left(\hat{\mathcal{A}}_k - \mathcal{A}_k\right) = \begin{pmatrix} X_k & Y_k \\ -\alpha\, Y_k & X_k \end{pmatrix}$$

with

$$\begin{aligned} X_k &:= \left(\alpha\hat{A}_k^{-1}\hat{B}_k + \hat{B}_k^{-1}\hat{A}_k\right)^{-1}\left(\alpha\hat{A}_k^{-1}\Delta B_k + \hat{B}_k^{-1}\Delta A_k\right) \\ Y_k &:= \left(\alpha\hat{A}_k^{-1}\hat{B}_k + \hat{B}_k^{-1}\hat{A}_k\right)^{-1}\left(\hat{B}_k^{-1}\Delta B_k - \hat{A}_k^{-1}\Delta A_k\right), \end{aligned}$$

where $\Delta A_k := \hat{A}_k - A_k$ and $\Delta B_k := \hat{B}_k - B_k$.

A similarity transformation with the matrix

$$\mathcal{N}_k := \begin{pmatrix} \mathrm{i}I & -\mathrm{i}I \\ \sqrt{\alpha}I & \sqrt{\alpha}I \end{pmatrix},$$

leads to a block-diagonal matrix $\mathcal{M}_k$:

$$\begin{aligned} \mathcal{M}_k &= \mathcal{N}_k^{-1}\left(I - \hat{\mathcal{A}}_k^{-1}\mathcal{A}_k\right)\mathcal{N}_k \\ &= \begin{pmatrix} X_k - \mathrm{i}\sqrt{\alpha}\, Y_k & \\ & X_k + \mathrm{i}\sqrt{\alpha}\, Y_k \end{pmatrix} \end{aligned}$$

with

$$\begin{aligned} X_k - \mathrm{i}\sqrt{\alpha}\, Y_k &= \left(\hat{A}_k + \sqrt{\alpha}\hat{B}_k\mathrm{i}\right)^{-1}\left(\Delta A_k + \sqrt{\alpha}\Delta B_k\mathrm{i}\right) \\ X_k + \mathrm{i}\sqrt{\alpha}\, Y_k &= \left(\hat{A}_k - \sqrt{\alpha}\hat{B}_k\mathrm{i}\right)^{-1}\left(\Delta A_k - \sqrt{\alpha}\Delta B_k\mathrm{i}\right). \end{aligned}$$

It is easy to see that

$$\mathcal{N}_k\mathcal{N}_k^H = 2\begin{pmatrix} I & \\ & \alpha\, I \end{pmatrix},$$

where $\mathcal{N}_k^H$ denotes the Hermitian transpose of $\mathcal{N}_k$. We introduce

$$\tilde{\mathcal{N}}_k := \frac{1}{2}\begin{pmatrix} (\hat{A}_k^2 + \alpha\hat{B}_k^2)^{-1/4} & \\ & (\hat{A}_k^2 + \alpha\hat{B}_k^2)^{-1/4} \end{pmatrix}\mathcal{N}_k$$

and obtain $\left(\tilde{\mathcal{N}}_k\tilde{\mathcal{N}}_k^H\right)^{-1} = \mathcal{L}_k$. The matrix

$$\tilde{\mathcal{M}}_k = \tilde{\mathcal{N}}_k^{-1}\left(I - \hat{\mathcal{A}}_k^{-1}\mathcal{A}_k\right)\tilde{\mathcal{N}}_k$$

is block diagonal with (1,1)-block $\tilde{Z}_k$. The (2,2)-block is the conjugate complex of the (1,1)-block. Therefore obviously

$$\left\|\tilde{\mathcal{M}}_k^\nu\right\|_{\ell^2} = \left\|\tilde{Z}_k^\nu\right\|_{\ell^2}$$

holds. Since

$$\begin{aligned}
\left\|\tilde{\mathcal{M}}_k^{\nu}\right\|_{\ell^2}^2 &= \left\|\left(\tilde{\mathcal{N}}_k^{-1}\left(I-\hat{\mathcal{A}}_k^{-1}\mathcal{A}_k\right)\tilde{\mathcal{N}}_k\right)^{\nu}\right\|_{\ell^2}^2 \\
&= \left\|\tilde{\mathcal{N}}_k^{-1}\left(I-\hat{\mathcal{A}}_k^{-1}\mathcal{A}_k\right)^{\nu}\tilde{\mathcal{N}}_k\right\|_{\ell^2}^2 \\
&= \rho\left(\tilde{\mathcal{N}}_k^{H}\left(I-\hat{\mathcal{A}}_k^{-1}\mathcal{A}_k\right)^{\nu T}\tilde{\mathcal{N}}_k^{-H}\tilde{\mathcal{N}}_k^{-1}\left(I-\hat{\mathcal{A}}_k^{-1}\mathcal{A}_k\right)^{\nu}\tilde{\mathcal{N}}_k\right) \\
&= \rho\left(\mathcal{L}_k^{-1}\left(I-\hat{\mathcal{A}}_k^{-1}\mathcal{A}_k\right)^{\nu T}\mathcal{L}_k\left(I-\hat{\mathcal{A}}_k^{-1}\mathcal{A}_k\right)^{\nu}\right) \\
&= \left\|\mathcal{L}_k^{1/2}\left(I-\hat{\mathcal{A}}_k^{-1}\mathcal{A}_k\right)^{\nu}\mathcal{L}_k^{-1/2}\right\|_{\ell^2}^2 \\
&= \left\|\left(I-\hat{\mathcal{A}}_k^{-1}\mathcal{A}_k\right)^{\nu}\right\|_{\mathcal{L}_k}^2,
\end{aligned}$$

the proof is completed. □

Lemma 54 *Under the assumptions and notations of Theorem 49 the matrix* $I-\hat{\mathcal{A}}_k^{-1}\mathcal{A}_k$ *is power bounded with constant* 2*, i.e., it satisfies* (4.34) *with* $C_B = 2$.

Proof: It is sufficient to show that $\tilde{Z}_k$, given in Lemma 53, is power bounded (with constant 2). We will show that

$$r(\tilde{Z}_k) \leq 1 \tag{4.35}$$

holds, where

$$r(\tilde{Z}_k) := \sup_{\underline{x}_k \in \mathbb{C}^{N_k}\setminus\{0\}} \left| \frac{(\tilde{Z}_k \underline{x}_k, \underline{x}_k)_{\ell^2}}{(\underline{x}_k, \underline{x}_k)_{\ell^2}} \right| \tag{4.36}$$

is the numerical radius of the matrix $\tilde{Z}_k$.

Observe that

$$Z_k = (a_k + \sqrt{\alpha} b_k \mathrm{i})^{-1} \hat{D}_k^{-1} (\Delta A_k + \sqrt{\alpha}\Delta B_k \mathrm{i})$$

and, therefore,

$$\begin{aligned}
\tilde{Z}_k &= (\hat{A}_k^2 + \alpha \hat{B}_k^2)^{1/4}\, Z_k\, (\hat{A}_k^2 + \alpha \hat{B}_k^2)^{-1/4} = \hat{D}_k^{1/2}\, Z_k\, \hat{D}_k^{-1/2} \\
&= (a_k + \sqrt{\alpha} b_k \mathrm{i})^{-1} \hat{D}_k^{-1/2} (\Delta A_k + \sqrt{\alpha}\Delta B_k \mathrm{i}) \hat{D}_k^{-1/2}.
\end{aligned}$$

Hence we obtain

$$\begin{aligned}
r(\tilde{Z}_k) &= \sup_{\underline{x}_k \in \mathbb{C}^{N_k}\setminus\{0\}} \left| \frac{(\tilde{Z}_k \underline{x}_k, \underline{x}_k)_{\ell^2}}{(\underline{x}_k, \underline{x}_k)_{\ell^2}} \right| \\
&= \sup_{\underline{x}_k \in \mathbb{C}^{N_k}\setminus\{0\}} \left| \frac{((\Delta A_k + \sqrt{\alpha}\Delta B_k \mathrm{i})\underline{x}_k, \underline{x}_k)_{\ell^2}}{(a_k + \sqrt{\alpha} b_k \mathrm{i})(\hat{D}_k^{1/2}\underline{x}_k, \hat{D}_k^{1/2}\underline{x}_k)_{\ell^2}} \right|
\end{aligned}$$

and further

$$r(\tilde{Z}_k) = \sup_{\underline{x}_k \in \mathbb{C}^{N_k} \setminus \{0\}} \left| \frac{(\Delta A_k \underline{x}_k, \underline{x}_k)_{\ell^2} + \sqrt{\alpha}(\Delta B_k \underline{x}_k, \underline{x}_k)_{\ell^2} \mathrm{i}}{(\hat{A}_k \underline{x}_k, \underline{x}_k)_{\ell^2} + \sqrt{\alpha}(\hat{B}_k \underline{x}_k, \underline{x}_k)_{\ell^2} \mathrm{i}} \right|$$
$$= \sup_{\underline{x}_k \in \mathbb{C}^{N_k} \setminus \{0\}} \sqrt{\frac{(\Delta A_k \underline{x}_k, \underline{x}_k)_{\ell^2}^2 + \alpha(\Delta B_k \underline{x}_k, \underline{x}_k)_{\ell^2}^2}{(\hat{A}_k \underline{x}_k, \underline{x}_k)_{\ell^2}^2 + \alpha(\hat{B}_k \underline{x}_k, \underline{x}_k)_{\ell^2}^2}}.$$

The last equation holds because all involved scalar products have real values. We know that numerical radius is bounded by 1, if we can show that $(\Delta A_k \underline{x}_k, \underline{x}_k)_{\ell^2}^2 \leq (\hat{A}_k \underline{x}_k, \underline{x}_k)_{\ell^2}^2$ and $(\Delta B_k \underline{x}_k, \underline{x}_k)_{\ell^2}^2 \leq (\hat{B}_k \underline{x}_k, \underline{x}_k)_{\ell^2}^2$ holds for all $\underline{x}_k \in \mathbb{C}^{N_k}$.

This property can be shown: The estimate (4.33) implies that

$$((\hat{A}_k^{-1/2} A_k \hat{A}_k^{-1/2} - I)\underline{x}_k, \underline{x}_k)_{\ell^2} \leq (\underline{x}_k, \underline{x}_k)_{\ell^2}$$

holds for all vectors $\underline{x}_k \in \mathbb{C}^{N_k}$, since $\hat{A}_k$ is symmetric and positive definite. Using $\Delta A_k = A_k - \hat{A}_k$, this implies

$$(\Delta A_k \underline{x}_k, \underline{x}_k)_{\ell^2} \leq (\hat{A}_k \underline{x}_k, \underline{x}_k)_{\ell^2}. \tag{4.37}$$

Since A_k is symmetric and positive definite, we have moreover

$$-(\hat{A}_k \underline{x}_k, \underline{x}_k)_{\ell^2} \leq (\Delta A_k \underline{x}_k, \underline{x}_k)_{\ell^2}. \tag{4.38}$$

Combining (4.37) and (4.38) shows that

$$(\Delta A_k \underline{x}_k, \underline{x}_k)_{\ell^2}^2 \leq (\hat{A}_k \underline{x}_k, \underline{x}_k)_{\ell^2}^2$$

holds for all $\underline{x}_k \in \mathbb{C}^{N_k}$. The argument for B_k is completely analogous.

Hence we have shown (4.35).

For the next step we use that the numerical radius satisfies the power inequality

$$r(M^\nu) \leq r(M)^\nu$$

for all (quadratic) matrices M and all $\nu \in \mathbb{N}$, see, e.g., Pearcy [46],.

Using the power inequality, the estimate (4.35) implies that

$$r\left(\tilde{Z}_k^\nu\right) \leq 1$$

holds for all $\nu \in \mathbb{N}$. Using the fact, that $\|M\|_{\ell^2} \leq 2\, r(M)$ holds for all matrices, we know that

$$\left\|\tilde{Z}_k^\nu\right\|_{\ell^2} \leq 2$$

holds for all $\nu \in \mathbb{N}$, which finishes the proof. □

Additionally, we need that the preconditioner $\hat{\mathcal{A}}_k$ can be bounded from above using the matrix $\mathcal{L}_k$:

Lemma 55 *Under the assumptions and notations of Theorem 49, we have*

$$\left\| \mathcal{L}_k^{-1/2} \hat{\mathcal{A}}_k \mathcal{L}_k^{-1/2} \right\|_{\ell^2} = 1.$$

Proof: Using the definition $\hat{Z}_k := \left(\hat{A}_k^2 + \alpha \hat{B}_k^2 \right)^{1/4}$, we observe that $\hat{Z}_k = \left(a_k^2 + \alpha b_k^2 \right)^{1/4} \hat{D}_k^{1/4}$. Therefore the desired result immediately follows:

$$\begin{aligned} \left\| \mathcal{L}_k^{-1/2} \hat{\mathcal{A}}_k \mathcal{L}_k^{-1/2} \right\|_{\ell^2} &= \left\| \begin{pmatrix} \hat{Z}_k^{-1} \hat{A}_k \hat{Z}_k^{-1} & \hat{Z}_k^{-1} \alpha^{1/2} \hat{B}_k \hat{Z}_k^{-1} \\ \hat{Z}_k^{-1} \alpha^{1/2} \hat{B}_k \hat{Z}_k^{-1} & -\hat{Z}_k^{-1} \hat{A}_k \hat{Z}_k^{-1} \end{pmatrix} \right\|_{\ell^2} \\ &= \left(a_k^2 + \alpha b_k^2 \right)^{-1/2} \left\| \begin{pmatrix} a_k I & \alpha^{1/2} b_k I \\ \alpha^{1/2} b_k I & -a_k I \end{pmatrix} \right\|_{\ell^2} = 1. \end{aligned}$$

□

We combine Lemmas 51 – 55 to prove Theorem 49 as follows.

Proof of Theorem 49: Let $\mathcal{M}_k = I - \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k$. Lemma 54 states that

$$\| \mathcal{M}_k^\nu \|_{\mathcal{L}_k} \leq 2,$$

i.e., condition (4.34), holds. Using Lemma 51 we conclude

$$\| (I - \mathcal{M}_k)((1-\tau)I + \tau \mathcal{M}_k)^\nu \|_{\mathcal{L}_k} \leq \frac{C}{\sqrt{\nu}}.$$

By plugging in for $\mathcal{M}_k$, we obtain

$$\left\| \mathcal{L}_k^{1/2} \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k \left(I - \tau \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k \right)^\nu \mathcal{L}_k^{-1/2} \right\|_{\ell^2} \leq \frac{C}{\sqrt{\nu}}.$$

Using the sub-multiplicativity of norms, we obtain

$$\left\| \mathcal{L}_k^{-1/2} \mathcal{A}_k \left(I - \tau \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k \right)^\nu \mathcal{L}_k^{-1/2} \right\|_{\ell^2} \leq \frac{C}{\sqrt{\nu}} \left\| \mathcal{L}_k^{-1/2} \hat{\mathcal{A}}_k \mathcal{L}_k^{-1/2} \right\|_{\ell^2},$$

which finishes the proof, as we know from Lemma 55 that $\left\| \mathcal{L}_k^{-1/2} \hat{\mathcal{A}}_k \mathcal{L}_k^{-1/2} \right\|_{\ell^2} = 1$. □

4.5 Application to the model problem 2: a robust convergence result based on partial regularity

The convergence proofs we have presented so far are based on the regularity assumption **(R)**, introduced on page 27, which cannot be guaranteed on domains with reentrant corners. One can show that on such domains there exist functions $f \in L^2(\Omega)$ such that the solution y_f of the problem, find $y_f \in H^1(\Omega)$ such that

$$(y_f, \tilde{y})_{H^1(\Omega)} = \langle f, \tilde{y} \rangle \qquad \text{for all } \tilde{y} \in H^1(\Omega),$$

has singularities close to the corners. This implies that y_f is not in $H^2(\Omega)$, but in a weaker space $H^{2-s}(\Omega)$ for some $s \in (0,1)$.

In this section, we carry out a convergence result that is based on the following regularity assumption.

(R') *Partial elliptic regularity:* There is a parameter $s \in (0,1)$ and a constant $C_R > 0$ such that the following result holds. For $f \in [H^s(\Omega)]^*$ let $y_f \in Y = H^1(\Omega)$ be such that

$$(y_f, \tilde{y})_{H^1(\Omega)} = \langle f, \tilde{y} \rangle \qquad \text{for all } \tilde{y} \in H^1(\Omega).$$

Then $y_f \in H^{2-s}(\Omega)$ and

$$\|y_f\|_{H^{2-s}(\Omega)} \leq C_R \|f\|_{[H^s(\Omega)]^*}.$$

The following theorem guarantees that the regularity assumption **(R')** is satisfied on general polygonal domains.

Theorem 56 *Let Ω be an open subset of $\mathbb{R}^2$ with polygonal boundary. The angels of the domain Ω, measured from inside, are denoted by ω_j for $j = 1, \ldots, M$. (So, $\omega_j > \pi$ refers to a reentrant corner.) Let ω be the largest angle, i.e., $\omega = \max_{j=1,\ldots,M}\{\omega_j\}$. Then the regularity assumption* **(R')** *holds for all s with*

$$\max\left\{0, 1 - \frac{\pi}{\omega}\right\} < s \leq 1.$$

For a proof see, e.g., Grisvard [33], Remark 2.4.6.

We can rewrite the lower bound for s in the following way: $s = \max\left\{0, 1 - \frac{\pi}{\omega}\right\} + \epsilon$ with $\epsilon > 0$. A simple example for a non-convex polygonal domain is the L-shaped domain (Figure 4.1). For the L-shaped domain, we obtain $s = \frac{1}{3} + \epsilon$.

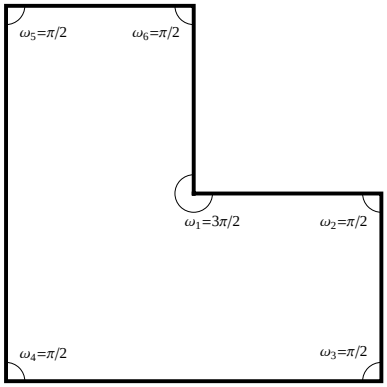


Figure 4.1: L-shaped domain

In the next subsections, we introduce a regularity result that is based on partial regularity. In Subsection 4.5.1 we discuss how to choose the norms for the case of partial regularity. In Subsections 4.5.2 and 4.5.3 we show the smoothing property and the approximation property for this choice of the norms. Finally, we summarize the results and give an overall convergence result. The results presented in this section have been published in the recent paper Takacs and Zulehner [63].

4.5.1 The choice of the norms

In this subsection we introduce the norms which we will use to carry out a convergence result assuming that regularity assumption **(R')** is guaranteed for some $s \in (0,1)$.

Here, the Hilbert spaces (their norms) $X^0_{-,k} := X_{-,k}$, X and $X^0_{+,k} := X_{+,k}$ are the Hilbert which we have introduced in Section 4.3 for showing the convergence for the full elliptic regularity case.

In this section, the convergence is shown in the Hilbert space $X^s_{-,k}$, given by interpolation:

$$X^s_{-,k} := [X_{-,k}, X]_s, \tag{4.39}$$

i.e., $X^s_{-,k}$ is the interpolation space of $X_{-,k}$ and X.

The next step is to construct a closed form of the norm on this Hilbert space. For this purpose, we need the following lemma.

Lemma 57 *For all Banach spaces A_1 and A_2, the Banach spaces $A_1 \cap [A_2, A_1]_\theta$ and $[A_1 \cap A_2, A_1]_\theta$ are equal and have equivalent norms. The constants, describing the equivalence, only depend on the choice of θ.*

Proof: In this proof, $C > 0$ is a generic constant which is independent of k and α but which may depend on θ. First note that

$$\|u\|_{[A_1\cap A_2,A_1]_\theta} \geq C\|u\|_{[A_1,A_1]_\theta\cap[A_2,A_1]_\theta} \geq C\|u\|_{A_1\cap[A_2,A_1]_\theta}$$

follows directly from the monotonicity of the interpolation.

So it remains to show $\|u\|_{[A_1\cap A_2,A_1]_\theta} \leq C\|u\|_{A_1\cap[A_2,A_1]_\theta}$. Let $u \in [A_1 \cap A_2]_\theta \cap A_1$. The definition of the norms on the interpolation spaces (real K-method, cf. Lions and Magenes [43]) and elementary relations yield

$$\begin{aligned}
\|u\|^2_{[A_1\cap A_2,A_1]_\theta} &= \int_0^\infty t^{-2\theta-1} \inf_{u_1+u_2=u} \left(\|u_1\|^2_{A_1} + \|u_1\|^2_{A_2} + t^2\|u_2\|^2_{A_1}\right) \,\mathrm{d}t \\
&= \int_0^1 t^{-2\theta-1} \inf_{u_1+u_2=u} \left(\|u_1\|^2_{A_1} + \|u_1\|^2_{A_2} + t^2\|u_2\|^2_{A_1}\right) \,\mathrm{d}t \\
&\quad + \int_1^\infty t^{-2\theta-1} \inf_{u_1+u_2=u} \left(\|u_1\|^2_{A_1} + \|u_1\|^2_{A_2} + t^2\|u_2\|^2_{A_1}\right) \,\mathrm{d}t.
\end{aligned}$$

By replacing the infimum by a particular choice, using the triangular inequality and by computing the integrals, we obtain

$$\begin{aligned}
&\|u\|^2_{[A_1\cap A_2,A_1]_\theta} \\
&\quad\leq \int_0^1 t^{-2\theta-1}t^2\|u\|^2_{A_1} \,\mathrm{d}t \\
&\qquad + \int_1^\infty t^{-2\theta-1} \inf_{u_1+u_2=u} \left((\|u\|_{A_1} + \|u_2\|_{A_1})^2 + \|u_1\|^2_{A_2} + t^2\|u_2\|^2_{A_1}\right) \,\mathrm{d}t \\
&\quad\leq \frac{1}{2-2\theta}\|u\|^2_{A_1} + \frac{1}{\theta}\|u\|^2_{A_1} + 2\int_1^\infty t^{-2\theta-1} \left(\inf_{u_1+u_2=u} \|u_1\|^2_{A_2} + (1+t)^2\|u_2\|^2_{A_1}\right) \,\mathrm{d}t.
\end{aligned} \tag{4.40}$$

By a variable transformation and again using the definition of the norms on the interpolation spaces, we obtain that further

$$\begin{aligned}
&\|u\|^2_{[A_1\cap A_2,A_1]_\theta} \\
&\quad\leq \frac{1}{(1-\theta)\theta}\|u\|^2_{A_1} \\
&\qquad + 2\left(\frac{1}{2}\right)^{-2\theta-1} \int_1^\infty (1+t)^{-2\theta-1} \inf_{u_1+u_2=u} \left(\|u_1\|^2_{A_2} + (1+t)^2\|u_2\|^2_{A_1}\right) \,\mathrm{d}t \\
&\quad= \frac{1}{(1-\theta)\theta}\|u\|^2_{A_1} + 2^{2\theta+2}\int_2^\infty t^{-2\theta-1} \inf_{u_1+u_2=u} \left(\|u_1\|^2_{A_2} + t^2\|u_2\|^2_{A_1}\right) \,\mathrm{d}t \\
&\quad\leq \frac{1}{(1-\theta)\theta}\|u\|^2_{A_1} + 2^{2\theta+2}\|u\|^2_{[A_2,A_1]_\theta} \leq C(\theta)^2\|u\|^2_{[A_2,A_1]_\theta\cap A_1}
\end{aligned}$$

holds, which finishes the proof for $C(\theta) = \max\left\{(1-\theta)^{-1/2}\theta^{-1/2}, 2^{\theta+1}\right\}$. □

For further reference, the following lemma gives a closed representation of the norm $\|\cdot\|_{X^s_{-,k}}$ or, more precisely, we introduce a norm with closed form which is equivalent to the norm $\|\cdot\|_{X^s_{-,k}}$.

Lemma 58 *The Hilbert space $X^s_{-,k}$, introduced in (4.39) by interpolation, is the linear space*

$$X^s_- = H^s(\Omega) \times H^s(\Omega)$$

equipped with the mesh-dependent norm

$$\|x\|_{X^s_{-,k}} = \left(\|y\|^2_{Y^s_{-,k}} + \|p\|^2_{P^s_{-,k}} \right)^{1/2},$$

where

$$\|y\|_{Y^s_{-,k}} \sim \left(1 + \alpha^{1/2} h_k^{-2}\right)^{(1-s)/2} \left(\|y\|^2_{L^2(\Omega)} + \alpha^{s/2} \|y\|^2_{H^s(\Omega)} \right)^{1/2} \text{ and}$$
$$\|p\|_{P^s_{-,k}} \sim \alpha^{-1} \left(1 + \alpha^{1/2} h_k^{-2}\right)^{(1-s)/2} \left(\|p\|^2_{L^2(\Omega)} + \alpha^{s/2} \|p\|^2_{H^s(\Omega)} \right)^{1/2}.$$

Here, $\sim$ denotes the equivalence of norms, where the constants are independent of h_k and α.

Proof: First note that $X^s_{-,k}$, defined by (4.39), has product structure. Therefore, it suffices to discuss the Hilbert spaces $Y^s_{-,k}$ and the $P^s_{-,k}$ separately. First, we consider $Y^s_{-,k}$.

Using (4.39), the definitions of the norms $\|\cdot\|_{Y^0_{-,k}}$ and $\|\cdot\|_Y$ and Lemma 57, we obtain

$$\begin{aligned}
\|y\|_{Y^s_{-,k}} &= \|y\|_{[(Y^0_{-,k}),Y]_s} \\
&= \|y\|_{[(1+\alpha^{1/2}h_k^{-2})^{1/2}L^2(\Omega), L^2(\Omega)\cap\alpha^{1/4}H^1(\Omega)]_s} \\
&\sim \left(1 + \alpha^{1/2} h_k^{-2}\right)^{(1-s)/2} (\|y\|^2_{L^2(\Omega)} + \alpha^{s/2} \|y\|_{H^s(\Omega)})^{1/2}.
\end{aligned}$$

The same can be done for the Hilbert space $P^s_{-,k}$. □

In the present section we show the convergence in the Hilbert space $X^s_{-,k}$. This is done using the smoothing property and the approximation property. In the context of the present section these properties read as follows.

- *Smoothing property:* There is some function η with $\lim_{\nu\to\infty} \eta(\nu) = 0$ such that for all grid levels $k \in \mathbb{N}$ and all $\nu \in \mathbb{N}$ the estimate

$$\sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\mathcal{B}\left(x_k^{(0,\nu)} - x_k, \tilde{x}_k\right)}{\|\tilde{x}_k\|_{X^s_{-,k}}} \leq \eta(\nu) \left\| x_k^{(0)} - x_k \right\|_{X^s_{-,k}} \tag{4.41}$$

  holds.

- *Approximation property:* There is a constant $C_A > 0$ such that for all grid levels $k \in \mathbb{N}$ the estimate

$$\left\| x_k^{(1)} - x_k \right\|_{X^s_{-,k}} \leq C_A \sup_{\tilde{x}_k \in X_k \setminus \{0\}} \frac{\mathcal{B}\left(x_k^{(0,\nu)} - x_k, \tilde{x}_k \right)}{\|\tilde{x}_k\|_{X^s_{-,k}}} \tag{4.42}$$

  holds.

Again, the combination of the smoothing property and the approximation property implies the convergence of the two-grid method.

4.5.2 Smoothing property

In this subsection we show that the smoothing property for the full elliptic regularity case can be carried over to the partial elliptic regularity case. Here, we are not interested in constructing a particular smoother for the partial regularity case. Therefore we do not show condition **(A2)**.

Note that it is not a good idea to set up a normal equation smoother in the Hilbert space $X^s_{-,k}$ because the a matrix representing the norm $\|\cdot\|_{X^s_{-,k}}$ cannot be inverted efficiently. We are interested in using *the same* smoothers that we have proposed for the full elliptic regularity case. The following theorem states that the smoothing property (4.41) is satisfied.

Theorem 59 *Assume that $\mathcal{A}_k$ is symmetric and that the smoother is given by*

$$\underline{x}_k^{(0,m)} := \underline{x}_k^{(0,m-1)} + \hat{\mathcal{A}}_k^{-1} \left(\underline{f}_k - \mathcal{A}_k \, \underline{x}_k^{(0,m-1)} \right) \qquad \text{for } m = 1, \ldots, \nu, \tag{4.43}$$

where $\hat{\mathcal{A}}_k$ is a symmetric matrix. Assume that this smoother satisfies the smoothing property in the norm $\|\cdot\|_{X^0_{-,k}}$, i.e.,

$$\sup_{\widetilde{x}_k \in X_k} \frac{\mathcal{B}\left(x_k^{(0,\nu)} - x_k, \widetilde{x}_k \right)}{\|\widetilde{x}_k\|_{X^0_{-,k}}} \leq \eta(\nu) \left\| x_k^{(0)} - x_k \right\|_{X^0_{-,k}}$$

holds. Moreover, assume that condition **(A1a)** *holds and the smoother is power-bounded, i.e., condition* (4.7) *holds.*

Then for all $s \in (0,1)$ the smoother satisfies the smoothing property also in the norm $\|\cdot\|_{X^s_{-,k}}$, i.e., there is a constant $\widetilde{C}_S$, depending only on s, $\underline{C}_D$ and $\overline{C}_D$ and C_B, such that

$$\sup_{\widetilde{x}_k \in X_k} \frac{\mathcal{B}\left(x_k^{(0,\nu)} - x_k, \widetilde{x}_k \right)}{\|\widetilde{x}_k\|_{X^s_{-,k}}} \leq \widetilde{C}_S \, \eta(\nu)^{1-s} \left\| x_k^{(0)} - x_k \right\|_{X^s_{-,k}} \tag{4.44}$$

is satisfied.

Proof: The proof is done using interpolation. By assumption, we know that the smoothing property

$$\sup_{\widetilde{x}_k \in X_k} \frac{\mathcal{B}\left(x_k^{(0,\nu)} - x_k, \widetilde{x}_k\right)}{\|\widetilde{x}_k\|_{X^0_{-,k}}} \leq \eta(\nu) \left\| x_k^{(0)} - x_k \right\|_{X^0_{-,k}}$$

is satisfied. We will also show that there is a constant $C > 0$ such that

$$\sup_{\widetilde{x}_k \in X_k} \frac{\mathcal{B}\left(x_k^{(0,\nu)} - x_k, \widetilde{x}_k\right)}{\|\widetilde{x}_k\|_X} \leq C \|x_k^{(0)} - x_k\|_X \tag{4.45}$$

holds. Then the interpolation theorem (Theorem 40) immediately implies (4.44).

In order to show (4.45), we reformulate the condition in matrix-vector notation:

$$\|\mathcal{A}_k (I - \tau \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k)^\nu \underline{r}_k\|_{\mathcal{Q}_k^{-1}} \leq C \|\underline{r}_k\|_{\mathcal{Q}_k}$$

has to be shown for all $\underline{r}_k := \underline{x}_k^{(0)} - \underline{x}_k$. Here, the matrix $\mathcal{Q}_k$ represents the scalar product $(\cdot,\cdot)_X$ on X_k. In other words, we have to show that the spectral norm of

$$\mathcal{P}_k := \mathcal{Q}_k^{-1/2} \mathcal{A}_k (I - \tau \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k)^\nu \mathcal{Q}_k^{-1/2}$$

is bounded by a constant. This matrix is symmetric, so we have

$$\begin{aligned} \|\mathcal{P}_k\|_{\ell^2} = \rho(\mathcal{P}_k) &= \rho\left(\mathcal{L}_k^{1/2} \mathcal{Q}_k^{-1} \mathcal{A}_k (I - \tau \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k)^\nu \mathcal{L}_k^{-1/2}\right) \\ &\leq \left\| \mathcal{L}_k^{1/2} \mathcal{Q}_k^{-1} \mathcal{A}_k \mathcal{L}_k^{-1/2} \right\|_{\ell^2} \left\| \mathcal{L}_k^{1/2} (I - \tau \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k)^\nu \mathcal{L}_k^{-1/2} \right\|_{\ell^2}, \end{aligned}$$

where $\|\cdot\|_{\ell^2}$ is the spectral norm and the matrix $\mathcal{L}_k$ represents the scalar product $(\cdot,\cdot)_{X^0_{-,k}}$ on X_k. Here, the second factor can be bounded from above by C_B using condition (4.7). The first factor can be bounded from above by two times the numerical radius, where $r(M)$ is the numerical radius of a matrix M. We obtain

$$\left\| \mathcal{L}_k^{1/2} \mathcal{Q}_k^{-1} \mathcal{A}_k \mathcal{L}_k^{-1/2} \right\|_{\ell^2} \leq 2\, r\left(\mathcal{L}_k^{1/2} \mathcal{Q}_k^{-1} \mathcal{A}_k \mathcal{L}_k^{-1/2}\right)$$

$$\begin{aligned} &\leq 2 \sup_{\underline{x}_k \in \mathbb{R}^{N_k} \setminus \{0\}} \frac{|(\mathcal{L}_k^{1/2} \mathcal{Q}_k^{-1} \mathcal{A}_k \mathcal{L}_k^{-1/2} \underline{x}_k, \underline{x}_k)_{\ell^2}|}{|(\underline{x}_k, \underline{x}_k)_{\ell^2}|} \\ &= 2 \sup_{\underline{x}_k \in \mathbb{R}^{N_k} \setminus \{0\}} \frac{|(\mathcal{Q}_k^{-1/2} \mathcal{L}_k \mathcal{Q}_k^{-1} \mathcal{A}_k \mathcal{Q}_k^{-1/2} \underline{x}_k, \underline{x}_k)_{\ell^2}|}{|(\mathcal{Q}_k^{-1/2} \mathcal{L}_k \mathcal{Q}_k^{-1/2} \underline{x}_k, \underline{x}_k)_{\ell^2}|} \end{aligned}$$

and further

$$\begin{aligned}
\left\|\mathcal{L}_k^{1/2}\mathcal{Q}_k^{-1}\mathcal{A}_k\mathcal{L}_k^{-1/2}\right\|_{\ell^2} &\leq 2 \sup_{\underline{x}_k,\underline{y}_k\in\mathbb{R}^{N_k}\setminus\{0\}} \frac{|(\mathcal{Q}_k^{-1/2}\mathcal{A}_k\mathcal{Q}_k^{-1/2}\underline{x}_k,\underline{y}_k)_{\ell^2}|}{|(\underline{x}_k,\underline{y}_k)_{\ell^2}|} \\
&\leq 2\left\|\mathcal{Q}_k^{-1/2}\mathcal{A}_k\mathcal{Q}_k^{-1/2}\right\|_{\ell^2} \\
&= 2 \sup_{\underline{x}_k\in\mathbb{R}^{N_k}\setminus\{0\}} \frac{|(\mathcal{Q}_k^{-1/2}\mathcal{A}_k\mathcal{Q}_k^{-1/2}\underline{x}_k,\underline{x}_k)_{\ell^2}|}{(\underline{x}_k,\underline{x}_k)_{\ell^2}} \\
&= 2 \sup_{\underline{x}_k\in\mathbb{R}^{N_k}\setminus\{0\}} \frac{|(\mathcal{A}_k\underline{x}_k,\underline{x}_k)_{\ell^2}|}{(\mathcal{Q}_k\underline{x}_k,\underline{x}_k)_{\ell^2}} \\
&= 2 \sup_{\underline{x}_k\in\mathbb{R}^{N_k}\setminus\{0\}} \frac{|\mathcal{B}(x_k,x_k)|}{\|x_k\|_X^2} \leq 2\overline{C}_D,
\end{aligned}$$

where $\overline{C}_D$ is the constant in **(A1a)**. This shows (4.45) which finishes the proof. □

The conditions of this theorem are satisfied for both smoothers introduced in this work, the collective Richardson smoother and the preconditioned normal equation smoother.

Corollary 60 *Both, the collective Richardson smoother and preconditioned normal equation smoother, satisfy the smoothing property* (4.41) *with smoothing rate*

$$\eta(\nu) = C_S(s)\,\nu^{-(1-s)/2},$$

where the constant $C_S(s)$ is independent of k and α. The constant $C_S(s)$ may depend on s.

Proof: For both smoothers, the smoothing property in the norm $\|\cdot\|_{X^0_{-,k}}$ was shown in Theorem 23 or Corollary 50, respectively. In both cases, the smoothing rate was given by $\eta(\nu) = C_S\nu^{-1/2}$.

Both methods can be represented in the closed from (4.43) with symmetric matrix $\hat{\mathcal{A}}_k$ and both methods are power-bounded (Lemma 24 or Lemma 54, respectively).

Therefore Theorem 59 implies the desired result. □

Remark 61 *We can show in a similar way that, provided power boundedness is satisfied for $s = 0$, that power boundedness is also satisfied for $s \in (0, 1)$.*

4.5.3 Approximation property

In this subsection, we show the approximation property by showing the conditions **(A1)**, **(A1a)**, **(A3)** and **(A4)**. Before we show the conditions, we have to introduce the Hilbert space $X^s_{+,k}$. The Hilbert spaces $X^s_{-,k}$ and X have already been introduced.

The Hilbert space $X^s_{+,k}$ is defined analogously to $X^s_{-,k}$ by interpolation:

$$X^s_{+,k} := [X_{+,k}, X]_s \tag{4.46}$$

For further reference, the following lemma gives a closed representation of the norm $\|\cdot\|_{X^s_{+,k}}$ or, more precisely, we introduce a norm with closed form which is equivalent to the norm $\|\cdot\|_{X^s_{+,k}}$.

Lemma 62 *The Hilbert space $X^s_{+,k}$, introduced in* (4.46)*, is the linear space*

$$X^s_+ = H^{2-s}(\Omega) \times H^{2-s}(\Omega)$$

equipped with a mesh-dependent norm

$$\|x\|_{X^s_{+,k}} = \left(\|y\|^2_{Y^s_{+,k}} + \|p\|^2_{P^s_{+,k}}\right)^{1/2},$$

where

$$\|y\|_{Y^s_{+,k}} \sim \left(1+\alpha^{1/2}h_k^{-2}\right)^{-(1-s)/2}\left(\|y\|^2_{L^2(\Omega)} + \alpha^{(2-s)/2}\|y\|^2_{H^{2-s}(\Omega)}\right)^{1/2} \text{ and}$$
$$\|p\|_{P^s_{+,k}} \sim \alpha^{-1}\left(1+\alpha^{1/2}h_k^{-2}\right)^{-(1-s)/2}\left(\|p\|^2_{L^2(\Omega)} + \alpha^{(2-s)/2}\|p\|^2_{H^{2-s}(\Omega)}\right)^{1/2}.$$

Here, $\sim$ denotes the equivalence of norms, where the constants are independent of h_k and α.

Proof: First note that $X^s_{+,k}$, defined by (4.46), has product structure. Therefore, it suffices to discuss the Hilbert spaces $Y^s_{+,k}$ and the $P^s_{+,k}$ separately. First, we consider $Y^s_{+,k}$.

Using (a) equation (4.46), (b) equation (4.29) and the reiteration theorem (Theorem 39), (c) the definitions of the norms $\|\cdot\|_{Y^0_{+,k}}$ and $\|\cdot\|_{Y^0_{-,k}}$ and (d) Lemma 57, we obtain

$$\begin{aligned}
\|y\|_{Y^s_{+,k}} &=_{\text{(a)}} \|y\|_{[Y_{+,k},Y]_s} \sim_{\text{(b)}} \|y\|_{[Y_{+,k},Y_{-,k}]_{s/2}} \\
&=_{\text{(c)}} \left(1+\alpha^{1/2}h_k^{-2}\right)^{-(1-s)/2} \|y\|_{[L^2(\Omega)\cap\alpha^{1/2}H^2(\Omega),L^2(\Omega)]_{s/2}} \\
&\sim_{\text{(d)}} \left(1+\alpha^{1/2}h_k^{-2}\right)^{(s-1)/2}\left(\|y\|^2_{L^2(\Omega)} + \alpha^{(2-s)/2}\|y\|^2_{H^{2-s}(\Omega)}\right)^{1/2}.
\end{aligned}$$

The same can be done for the Hilbert space $P^s_{+,k}$. □

The next step is to show the conditions **(A1)**, **(A1a)**, **(A3)** and **(A4)**.

Conditions (A1) and (A1a)

The conditions **(A1)** and **(A1a)** have already be shown in Theorem 12.

Condition (A3)

In Lemma 43, we have shown that the condition **(A3)**, introduced on page 56, is satisfied for $s = 0$, i.e., we have shown that

$$\inf_{x_k \in X_k} \|x - x_k\|_X \leq C_I \|x\|_{X^0_{+,k}}$$

holds for all $x \in X^0_+$.

Based on this result, the following lemma states that condition **(A3)** is also satisfied for $s \in (0,1)$.

Lemma 63 *In the framework of this section, condition* **(A3)** *is satisfied for all* $s \in (0,1)$*, i.e.,*

$$\inf_{x_k \in X_k} \|x - x_k\|_X \leq C_I \|x\|_{X^s_{+,k}}$$

holds for all $x \in X^s_+$.

Proof: Here, the analysis for the state y and for the adjoined state p completely decouples. We consider the state y first. We know that condition **(A3)** holds for $s = 0$, which implies that

$$\|y - \Pi_k y\|_Y \leq C \|y\|_{Y^0_{+,k}} \tag{4.47}$$

holds for all $y \in Y^0_+$.

Due to Theorem 16, we know that there is a projection operator Π_k on $Y = H^1(\Omega)$ such that the estimate (4.47) and the following boundedness result hold.

$$\|y - \Pi_k y\|_{H^1(\Omega)} \leq C \|y\|_{H^1(\Omega)} \qquad \text{and} \qquad \|y - \Pi_k y\|_{L^2(\Omega)} \leq C \|y\|_{L^2(\Omega)}.$$

This implies that

$$\|y - \Pi_k y\|_Y \leq C \|y\|_Y$$

holds for all $y \in Y$. The interpolation theorem (Theorem 40), relation (4.24) and the fact that $Y + Y_+^0 = Y$ states that

$$\sqrt{\frac{1}{2s(1-s)}}\|y - \Pi_k y\|_Y = \|y - \Pi_k y\|_{[Y,Y]_s} \leq C\|y\|_{[(Y_{+,k}^0),Y]_s} = C\|y\|_{Y_{+,k}^s}$$

holds for all $y \in Y_+^s$. The analysis for p is completely analogous. □

Condition (A4)

As it was done in Section 4.5, condition **(A4)**, introduced on page 56, is shown in two steps.

First note that one can show completely analogous to Lemma 30, that for $\mathcal{F} \in (X_-)^* = (H^s(\Omega))^* \times (H^s(\Omega))^*$, the corresponding solution $x_{\mathcal{F}} \in X_+ = H^{2-s}(\Omega) \times H^{2-s}(\Omega)$.

This does not allow to construct an estimate which is robust in α and k. For constructing an estimate, we use Remark 27 and show that the inf-sup-condition (4.12) is satisfied. Theorem 41 states that condition (4.12) is a consequence of conditions **(A4')** and **(A4")**.

First, we show the condition **(A4')**. Using the closed forms of the norms, introduced in Lemma 58 and Lemma 62, this condition reads as follows:

$$\begin{aligned} 0 \leq (y,y)_{L^2(\Omega)}^{1/2} &\leq C_{R3}\psi_k^{-1}\left(1+\alpha^{1/2}h_k^{-2}\right)^{(1-s)/2}\left(\|y\|_{L^2(\Omega)}^2 + \alpha^{s/2}\|y\|_{H^s(\Omega)}^2\right)^{1/2} \\ &\leq C_{R4}\psi_k\left(1+\alpha^{1/2}h_k^{-2}\right)^{-(1-s)/2}\left(\|y\|_{L^2(\Omega)}^2 + \alpha^{(2-s)/2}\|y\|_{H^{2-s}(\Omega)}^2\right)^{1/2} \end{aligned}$$

and

$$\begin{aligned} 0 \leq \alpha^{-1/2}(p,p)_{L^2(\Omega)}^{1/2} &\leq C_{R3}\alpha^{-1/2}\psi_k^{-1}\left(1+\alpha^{1/2}h_k^{-2}\right)^{(1-s)/2}\left(\|p\|_{L^2(\Omega)}^2 + \alpha^{s/2}\|p\|_{H^s(\Omega)}^2\right)^{1/2} \\ &\leq C_{R4}\alpha^{-1/2}\psi_k\left(1+\alpha^{1/2}h_k^{-2}\right)^{-(1-s)/2}\left(\|p\|_{L^2(\Omega)}^2 + \alpha^{(2-s)/2}\|p\|_{H^{2-s}(\Omega)}^2\right)^{1/2}. \end{aligned}$$

This is satisfied for $C_{R3} = C_{R4} = 1$ and $\psi_k := (1+\alpha^{1/2}h_k^{-2})^{(1-s)/2}$ because $\|\cdot\|_{H^{2-s}(\Omega)} \geq \|\cdot\|_{H^s(\Omega)}$ holds.

So, it remains to show the condition **(A4")**, i.e., that

$$C_{R1}\|y\|_{Y_{+,k}^s} \leq \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega)}}{\|\tilde{y}\|_{Y_{-,k}^s}} + \sup_{\tilde{p}\in P\setminus\{0\}} \frac{(y,\tilde{p})_{H^1(\Omega)}}{\|\tilde{p}\|_{P_{-,k}^s}} \tag{4.48}$$

holds for all $y \in Y_+^s$. By plugging in $\|\cdot\|_{P_{-,k}^s} = \alpha^{-1/2}\|\cdot\|_{Y_{-,k}^s}$ and combining the two suprema to one supremum, we obtain that

$$C_{R1}\|y\|_{Y_{+,k}^s} \leq \sup_{\tilde{y}\in Y\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega)} + \alpha^{1/2}(y,\tilde{p})_{H^1(\Omega)}}{\|\tilde{y}\|_{Y_{-,k}^s}} \qquad \text{for all } y \in Y_+^s \tag{4.49}$$

implies (4.48). Using the definition of the norms, (4.49) reads as follows:

$$C_{R1}\|y\|_{L^2(\Omega)\cap\alpha^{1/2-s/4}H^{2-s}(\Omega)} \leq \sup_{\tilde{y}\in H^1(\Omega)\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega)} + \alpha^{1/2}(y,\tilde{y})_{H^1(\Omega)}}{\|\tilde{y}\|_{L^2(\Omega)\cap\alpha^{s/4}H^s(\Omega)}} \tag{4.50}$$

holds for all $y \in H^{2-s}(\Omega)$.

For showing (4.50), we analyze the elliptic problem (4.51) first.

Lemma 64 *Assume that the assumption* **(R')** *is satisfied for some* $s \in (0,1)$.

Then there is a constant $C_E > 0$ *such that for all* $\alpha > 0$ *and all* $f \in [H^s(\Omega)]^*$ *the solution of the problem, find* $y_f \in H^1(\Omega)$ *such that*

$$(y_f,\tilde{y})_{L^2(\Omega)} + \alpha^{1/2}(y_f,\tilde{y})_{H^1(\Omega)} = \langle f,\tilde{y}\rangle \text{ for all } \tilde{y} \in H^1(\Omega), \tag{4.51}$$

satisfies $y_f \in H^{2-s}(\Omega)$ *and*

$$\|y_f\|_{L^2(\Omega)\cap\alpha^{1/2-s/4}H^{2-s}(\Omega)} \leq C_E\|f\|_{[L^2(\Omega)]^*+\alpha^{s/4}[H^s(\Omega)]^*}$$

holds. The constant C_E *only depends on the constant in the assumption* **(R')**.

Proof: Let $f \in [H^s(\Omega)]^*$. The first step is to show that $y_f \in H^{2-s}(\Omega)$.

From $y_f \in H^1(\Omega)$ we conclude $(y_f,\cdot)_{L^2(\Omega)} \in [H^s(\Omega)]^*$. Consider the following problem. Find $y \in Y$ such that

$$(y,\tilde{y})_{H^1(\Omega)} = \left\langle \alpha^{-1/2}(f-y_f),\tilde{y}\right\rangle \qquad \text{holds for all } \tilde{y} \in H^1(\Omega).$$

The regularity assumption **(R')** states $y_f \in H^{2-s}(\Omega)$ and

$$\|y_f\|_{H^{2-s}(\Omega)} \leq C_R\left(\|f\|_{[H^s(\Omega)]^*} + \alpha^{-1/2}\|y_f\|_{[H^s(\Omega)]^*}\right). \tag{4.52}$$

Condition **(A1)** implies $\|y_f\|_{H^1(\Omega)} \leq \underline{C}^{-1}\|f\|_{[H^s(\Omega)]^*}$. By combining this estimate with (4.52), we obtain

$$\|y_f\|_{H^{2-s}(\Omega)} \leq C(\alpha)\|f\|_{[H^s(\Omega)]^*},$$

where $C(\alpha)$ is some constant that may depend on α.

Now, in a second step we construct a result that is robust in α. Let $f \in L^2(\Omega)$ be arbitrarily but fixed.

We consider the following problem with solution y_f: Find $y \in H^1(\Omega)$ such that

$$(y, \tilde{y})_{H^1(\Omega)} = \left(\alpha^{-1/2}(f - y_f), \tilde{y}\right)_{L^2(\Omega)} \qquad \text{holds for all } \tilde{y} \in H^1(\Omega).$$

The Lax-Milgram theorem (Theorem 4) applied directly to the energy norm $\|\cdot\|_{H^1(\Omega)}$ shows that the solution $y_f \in H^1(\Omega)$ satisfies

$$\|y_f\|_{H^1(\Omega)} = \alpha^{-1/2}\|f - y_f\|_{[H^1(\Omega)]^*} \tag{4.53}$$

and the regularity assumption **(R')** implies that $y_f \in H^{2-s}(\Omega)$ and

$$\|y_f\|_{H^{2-s}(\Omega)} \leq C_R\, \alpha^{-1/2}\|f - y_f\|_{[H^s(\Omega)]^*}. \tag{4.54}$$

We consider the following problem with solution y_f: Find $y \in H^1(\Omega)$ such that

$$(y, \tilde{y})_{L^2(\Omega)} + \alpha^{1/2}(y, \tilde{y})_{H^1(\Omega)} = (f, \tilde{y})_{L^2(\Omega)} \qquad \text{holds for all } \tilde{y} \in H^1(\Omega). \tag{4.55}$$

The Lax-Milgram theorem applied directly to the energy norm $\|\cdot\|_{L^2(\Omega)\cap\alpha^{1/4}H^1(\Omega)}$ shows that the solution y_f satisfies

$$\|y_f\|_{L^2(\Omega)\cap\alpha^{1/4}H^1(\Omega)} = \|f\|_{[L^2(\Omega)\cap\alpha^{1/4}H^1(\Omega)]^*}. \tag{4.56}$$

The combination of (4.53) and (4.56) shows:

$$\|f - y_f\|_{[\alpha^{1/4}H^1(\Omega)]^*} \leq C\|f\|_{[L^2(\Omega)\cap\alpha^{1/4}H^1(\Omega)]^*}. \tag{4.57}$$

If we choose $\tilde{y} = y = y_f$ in (4.55), we obtain using $\alpha^{1/2}(y_f, \tilde{y})_{H^1(\Omega)} \geq 0$ that $\|y_f\|^2_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|y_f\|_{L^2(\Omega)}$ and therefore

$$\|y_f\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}$$

and therefore

$$\|f - y_f\|_{L^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}. \tag{4.58}$$

The combination of (4.57) and (4.58) and the interpolation theorem (Theorem 40) together with (4.26) shows:

$$\|f - y_f\|_{[\alpha^{s/4}H^s(\Omega)]^*} \leq C\|f\|_{[L^2(\Omega)\cap\alpha^{s/4}H^s(\Omega)]^*},$$

which reads, if combined with (4.54), as follows:

$$\|y_f\|_{\alpha^{1/2-s/4}H^{2-s}(\Omega)} \leq C\|f\|_{[L^2(\Omega)\cap\alpha^{s/4}H^s(\Omega)]^*}. \tag{4.59}$$

The equation (4.56) implies

$$\|y_f\|_{L^2(\Omega)} \leq \|f\|_{[L^2(\Omega)\cap\alpha^{1/4}H^1(\Omega)]^*},$$

which shows using (4.24)

$$\|y_f\|_{L^2(\Omega)} \leq \|f\|_{[L^2(\Omega)\cap\alpha^{1/4}H^s(\Omega)]^*},$$

which can be combined with (4.59) to the desired result:

$$\|y_f\|_{L^2(\Omega)\cap\alpha^{1/2-s/4}H^{2-s}(\Omega)} \leq C\|f\|_{L^2(\Omega)+\alpha^{s/4}[H^s(\Omega)]^*}$$

for all $f \in L^2(\Omega)$.

Since $L^2(\Omega)$ is dense in $[H^s(\Omega)]^*$ we have for $f_0 \in [H^s(\Omega)]^*$ and $f_\epsilon \in L^2(\Omega)$ with $\|f_0 - f_\epsilon\|_{[H^s(\Omega)]^*} \leq \epsilon$ that

$$\begin{aligned}
\|y_{f_0}\|_{L^2(\Omega)\cap\alpha^{1/2-s/4}H^{2-s}(\Omega)} &\leq \|y_{f_\epsilon}\|_{L^2(\Omega)\cap\alpha^{1/2-s/4}H^{2-s}(\Omega)} + \|y_{f_\epsilon} - y_{f_0}\|_{L^2(\Omega)\cap\alpha^{1/2-s/4}H^{2-s}(\Omega)} \\
&\leq C\|f_\epsilon\|_{L^2(\Omega)+\alpha^{s/4}[H^s(\Omega)]^*} + C(\alpha)\|f_\epsilon - f_0\|_{[H^s(\Omega)]^*} \\
&\leq C\|f_0\|_{L^2(\Omega)+\alpha^{s/4}[H^s(\Omega)]^*} + (1 + C(\alpha))\|f_\epsilon - f_0\|_{[H^s(\Omega)]^*} \\
&\leq C\|f_0\|_{L^2(\Omega)+\alpha^{s/4}[H^s(\Omega)]^*} + (1 + C(\alpha))\epsilon
\end{aligned}$$

holds, which shows the desired result for $\epsilon \to 0$. Here, y_{f_0} and y_{f_ϵ} are the solutions of the variational problem (4.51) for right-hand-sides f_0 and f_ϵ, respectively. □

Using the fact that $H^1(\Omega)$ is dense in $H^s(\Omega)$, the statement of Lemma 64 implies

$$\begin{aligned}
\sup_{\tilde{y}\in H^1(\Omega)\setminus\{0\}} \frac{(y,\tilde{y})_{L^2(\Omega)} + \alpha^{1/2}(y,\tilde{y})_{H^1(\Omega)}}{\|\tilde{y}\|_{L^2(\Omega)\cap\alpha^{s/4}H^s(\Omega)}} &= \sup_{\tilde{y}\in H^1(\Omega)\setminus\{0\}} \frac{\langle f,\tilde{y}\rangle}{\|\tilde{y}\|_{L^2(\Omega)\cap\alpha^{s/4}H^s(\Omega)}} \\
&= \sup_{\tilde{y}\in H^s(\Omega)\setminus\{0\}} \frac{\langle f,\tilde{y}\rangle}{\|\tilde{y}\|_{L^2(\Omega)\cap\alpha^{s/4}H^s(\Omega)}} \\
&= \|f\|_{[L^2(\Omega)\cap\alpha^{s/4}H^s(\Omega)]^*} \\
&\geq C_E^{-1}\|y\|_{L^2(\Omega)\cap\alpha^{1/2-s/4}H^{2-s}(\Omega)}.
\end{aligned}$$

This shows (4.50) and therefore **(A4")**.

As we have shown **(A4')** and **(A4")**, Theorem 41 implies (4.12) and, as a consequence, **(A4)**.

Approximation property

As we have shown **(A1)**, **(A1a)**, **(A3)** and **(A4)**, we can apply Theorem 26 and conclude as follows.

Corollary 65 *Consider Model Problem 2, assume that the regularity assumption* **(R')** *is satisfied for some $s \in (0,1)$. Then the approximation property holds with a constant C_A independent of the grid level and the choice of α.*

4.5.4 Convergence result

Again, the combination of approximation property and smoothing property shows the convergence of the two-grid method.

We could show that the preconditioned normal equation smoother and the collective Richardson smoother satisfy the conditions of the last subsections, i.e., we could show that the smoothing property for these methods holds. This shows – if also the conditions of Corollary 65 are satisfied – that the two-grid method converges.

For showing that also the W-cycle multigrid method converges, we have to show condition **(A5)**. This condition was satisfied for the case $s = 0$, i.e., we had

$$\underline{C}_C \|x_{k-1}\|_{X^0_{-,k}} \leq \|x_{k-1}\|_{X^0_{-,k-1}} \leq \overline{C}_C \|x_{k-1}\|_{X^0_{-,k}}$$

for all $x_{k-1} \in X_{k-1}$. Of course, also

$$\|x_{k-1}\|_X \leq \|x_{k-1}\|_X \leq \|x_{k-1}\|_X$$

is satisfied. Using the interpolation theorem (Theorem 40) we obtain

$$\underline{\hat{C}}_C \|x_{k-1}\|_{[(X^0_{-,k}),X]_s} \leq \|x_{k-1}\|_{[(X^0_{-,k-1}),X]_s} \leq \overline{\hat{C}}_C \|x_{k-1}\|_{[(X^0_{-,k}),X]_s},$$

and further

$$\underline{\hat{C}}_C \|x_{k-1}\|_{X^s_{-,k}} \leq \|x_{k-1}\|_{X^s_{-,k-1}} \leq \overline{\hat{C}}_C \|x_{k-1}\|_{X^s_{-,k}}.$$

Using the smoothing property (Corollary 60), the approximation property (Corollary 65) and condition **(A5)** we conclude as follows.

Corollary 66 *Consider Model Problem 2, assume that regularity assumption* **(R')** *is satisfied for some $s \in (0,1)$. Assume that the normal equation smoother (Subsections 3.2.1 and 4.1.4) or that the collective Richardson smoother is applied.*

Then there is a constant $C > 0$ independent of the grid level k and the choice of the parameter α such that

$$\left\| x_k^{(1)} - x_k \right\|_{X^s_{-,k}} \leq \frac{C}{\nu^{(1-s)/2}} \left\| x_k^{(0)} - x_k \right\|_{X^s_{-,k}}$$

holds, where x_k is the exact solution, $x_k^{(0)}$ is the starting value and $x_k^{(1)}$ is the iterate after one step of the two-grid or the W-cycle multigrid method.

Therefore, for ν large enough, the convergence rate is bounded away from 1 by a constant independent of the grid level k and the choice of α. The convergence rate may depend on s.

4.6 Summary

We could show for *all model problems* and for both, the reduced KKT-system and the non-reduced KKT-system, that the W-cycle multigrid iteration scheme with the *preconditioned normal equation smoother* converges if ν is large enough, i.e., we obtain a that there is a convergence rate $q \in (0, 1)$ (independent of the grid level k) such that

$$\left\| x_k^{(n)} - x_k \right\|_{L^2(\Omega)} \leq q^n \|y_D\|_{L^2(\Omega)}$$

holds on all grid levels, provided $x_k^{(0)} = 0$. Here, x_k is the exact solution and $x_k^{(n)}$ is the n-th iterate.

For the reduced KKT-system for *Model Problem 2*, we have shown convergence for two smoothers: the preconditioned normal equation smoother and the collective Richardson smoother. We have shown in Sections 4.3 and 4.4 that we have in both cases the same result as above but with convergence rate q independent of the grid level and of the choice of the parameter α. In Section 4.5, we have relaxed the regularity assumption **(R)** to regularity assumption **(R')** which does not exclude reentrant corners.

Chapter 5

Local Fourier analysis

Local Fourier analysis (or local mode analysis) is a commonly used approach for designing and analyzing convergence properties of multigrid methods. In the late 1970s A. Brandt proposed to use Fourier series to analyze multigrid methods, see, e.g., Brandt [19]. Local Fourier analysis provides a framework to analyze various numerical methods with a unified approach that gives quantitative statements on the methods under investigation. The computed bounds for the convergence rates are typically sharp. Other work on multigrid theory – such as the analysis presented in the last chapter – typically just shows convergence and does not give sharp or realistic bounds for the convergence rates.

Local Fourier analysis can be justified rigorously only in special cases, e.g., on rectangular domains with uniform grids and periodic boundary conditions, see, e.g., Brandt [20]. However, local Fourier analysis can also be interpreted as an heuristic approach for a wide class of applications.

For the analysis of multigrid methods for saddle point problems, local Fourier analysis has been applied recently, e.g., in Trottenberg [66], Borzi, Kunisch and Kwak [12] and Lass [41]. In Wienands [68] the method is explained as machinery and a local Fourier analysis software *LFA* is presented. This software can be configured using a graphical user interface and allows to approximate (numerically) smoothing and convergence rates based on local Fourier analysis approaches for various problems and multigrid approaches.

Upper bounds for smoothing rates or convergence rates can be formulated in terms of logical formulas consisting of quantifiers and polynomial inequalities. These formulas can be simplified by means of quantifier elimination using cylindrical algebraic decomposition. This tool has been applied earlier for finite difference methods for (systems of) ordinary and partial differential equations. There Hong, Liska and Steinberg [39]

have transformed the necessary conditions for stability, asymptotic stability and well-posedness of the given systems into statements on polynomial inequalities using Fourier or Laplace transforms.

For applying local Fourier analysis, we have to restrict ourselves to Model Problem 2. We consider the 2-by-2 formulation (reduced KKT-system) of the problem and the analysis is done for collective point smoothers (collective Jacobi and collective Gauss-Seidel smoother). Note that the presented strategy for the computation of the convergence rates is not restricted to this choice of the formulation of the problem or to the particular smoother.

The results presented in this chapter were worked out in a joint work with V. Pillwein. The analysis for the one dimensional case was partly published in Pillwein and Takacs [48]. For the two dimensional case, see Pillwein and Takacs [49], which is not a part of the thesis. The authors of that paper have prepared Mathematica notebooks[1] which contain the computations presented in these papers.

This chapter is organized as follows. The local Fourier analysis framework is introduced in Section 5.1. In Section 5.2, we will give a brief overview on quantifier elimination and cylindrical algebraic decomposition, i.e., on the symbolic methods applied in order to compute suprema symbolically. In Sections 5.3 and 5.4 we will apply the machinery introduced in the first two sections to the model problem.

5.1 Local Fourier analysis framework

5.1.1 Iteration matrix

Here, we restrict ourselves to the two-grid analysis: We consider a two-grid iteration scheme with $\nu_{pre} = \nu/2$ pre-smoothing and $\nu_{post} = \nu/2$ post-smoothing steps.

The main goal of a convergence analysis is to find a (sharp) bound for the convergence rate. This bound is the smallest factor q such that the norm of the error after the $n+1$-st iterate can be bounded by q times the error after the n-th iterate, i.e., such that

$$\left\| \underline{x}_k^{(n+1)} - \underline{x}_k \right\|_X \leq q \left\| \underline{x}_k^{(n)} - \underline{x}_k \right\|_X$$

is satisfied, where $\underline{x}_k := \mathcal{A}_k^{-1} \underline{f}_k$ is the exact solution and $\| \cdot \|_X$ is a given norm, as we will discuss later.

[1] The document is available online at http://www.risc.jku.at/people/vpillwei/sLFA/

Using the notations of Chapter 3, we obtain

$$\underline{x}_k^{(n+1)} - \underline{x}_k = TG_k^{k-1}\left(\underline{x}_k^{(n)} - \underline{x}_k\right),$$

where the iteration matrix TG_k^{k-1} is given by

$$TG_k^{k-1} := S_k^{\nu_{post}} \underbrace{\left(I - I_{k-1}^k \mathcal{A}_{k-1}^{-1} I_k^{k-1} \mathcal{A}_k\right)}_{CG_k^{k-1} :=} S_k^{\nu_{pre}},$$

and the iteration matrix of the smoother, S_k, is given by

$$S_k := I - \tau \hat{\mathcal{A}}_k^{-1} \mathcal{A}_k.$$

Certainly, the convergence rate can be bounded from above by the matrix norm of the iteration matrix, i.e.,

$$q \leq q_{TG} = \left\| TG_k^{k-1} \right\|_X$$

holds. This estimate is sharp if we consider the supremum over all possible starting values or, equivalently, all possible right-hand sides. If $q_{TG} < 1$ is satisfied, the method converges for all starting values with a contraction rate bounded by q_{TG}.

5.1.2 Symbols of the mass matrix and the stiffness matrix

The idea of (local) Fourier analysis is to simplify the problem such that the eigenvectors and the eigenvalues of the mass matrix and the stiffness matrix can be written down explicitly. Therefore, typically uniform grids are assumed. Whereas more rigorous approaches assume the domain to be an interval or a rectangle, in local Fourier analysis the boundary is neglected by assuming periodic boundary conditions. This allows to extend a bounded domain Ω to the entire space $\mathbb{R}$, see Brandt [20]. So we consider the case $\Omega := \mathbb{R}$. Let us repeat that the fact that local Fourier analysis predicts good convergence rates for simple cases, typically also indicates good convergence behavior of the analyzed methods on more general domains, cf. Figure 6.6 in Section 6.1.

As mentioned in Chapter 2, the discretization is done using the Courant element. Here, we use this fact directly. So, on each grid level $k = 0, 1, 2, \ldots$, we assume to have a uniform grid with nodes

$$x_{k,n} := n\, h_k \qquad \text{for } n \in \mathbb{Z},$$

where the uniform grid size is given by $h_k = 2^{-k}$. The functions in $Y_k = P_k$ are continuous on the whole domain and linear between two nodes. Therefore, the discretized function can be specified by prescribing the values on the nodes only.

For every $\theta \in \Theta := [-\pi, \pi)$ and every grid level k, we define a Fourier vector $\underline{\varphi_k}(\theta) \in \mathbb{C}^{\mathbb{Z}}$ as follows:

$$\underline{\varphi_k(\theta)} := (\varphi_{k,n}(\theta))_{n\in\mathbb{Z}} := (\mathrm{e}^{\mathrm{i}\theta x_{k,n}/h_k})_{n\in\mathbb{Z}}.$$

The next step is to analyze how the multiplication of the mass matrix and the stiffness matrix with the Fourier vectors looks like.

First, note that for uniform grids, the mass matrix and the stiffness matrix look as follows:

$$M_k = \frac{h_k}{6}\begin{pmatrix} \ddots & \ddots & \ddots & & & \\ & 1 & 4 & 1 & & \\ & & 1 & 4 & 1 & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}$$

and

$$K_k = \frac{1}{h_k}\begin{pmatrix} \ddots & \ddots & \ddots & & & \\ & -1 & 2 & -1 & & \\ & & -1 & 2 & -1 & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}.$$

If we consider an infinite domain, M_k and K_k become operators $\mathbb{C}^{\mathbb{Z}} \to \mathbb{C}^{\mathbb{Z}}$, given by

$$M_k\ \underline{\varphi_k(\theta)} = \left(\frac{h_k\,\varphi_{k,n-1}(\theta)}{6} + \frac{4\,h_k\,\varphi_{k,n}(\theta)}{6} + \frac{h_k\,\varphi_{k,n+1}(\theta)}{6}\right)_{n\in\mathbb{Z}}$$

and

$$K_k\ \underline{\varphi_k(\theta)} = \left(-\frac{\varphi_{k,n-1}(\theta)}{h_k} + \frac{2\,\varphi_{k,n}(\theta)}{h_k} - \frac{\varphi_{k,n+1}(\theta)}{h_k}\right)_{n\in\mathbb{Z}}.$$

We obtain

$$M_k\ \underline{\varphi_k(\theta)} = \underbrace{\frac{(e^{-i\theta}+4+e^{i\theta})h_k}{6}}_{\overline{M_k}(\theta) :=}\ \underline{\varphi_k(\theta)} \quad \text{and} \quad K_k\ \underline{\varphi_k(\theta)} = \underbrace{\frac{-e^{-i\theta}+2-e^{i\theta}}{h_k}}_{\overline{K_k}(\theta) :=}\ \underline{\varphi_k(\theta)}.$$

Thus indeed the Fourier vectors $\underline{\varphi_k(\theta)}$ are eigenvectors of M_k and K_k with eigenvalues $\overline{M_k}(\theta)$ and $\overline{K_k}(\theta)$, respectively. The quantities $\overline{M_k}(\theta)$ and $\overline{K_k}(\theta)$ are called symbols.

If we assume that $\underline{y_k}$ can be represented as a linear combination of Fourier vectors, i.e.,

$$\underline{y_k} = \sum_i \overline{y_k}(\theta_i)\ \underline{\varphi_k(\theta_i)} \tag{5.1}$$

we know that also the products $M_k \underline{y}_k$ and $K_k \underline{y}_k$ can be represented as linear combination of Fourier vectors:

$$M_k \underline{y}_k = \sum_i \overline{M_k}(\theta_i)\, \overline{y_k}(\theta_i)\, \underline{\varphi_k(\theta_i)} \qquad \text{and} \qquad K_k \underline{y}_k = \sum_i \overline{K_k}(\theta_i)\, \overline{y_k}(\theta_i)\, \underline{\varphi_k(\theta_i)}.$$

For bounded domains, it is possible to show that the decomposition (5.1) exists, cf. Brandt [20], Section 8. Often the existence of such a decomposition is just assumed, cf. Brandt [20], Section 3.1. In this case, local Fourier analysis is a formal tool. In the present work, we follow this idea.

5.1.3 Symbol of the system matrix $\mathcal{A}_k$

In the last subsection we have shown that for all $\theta \in \Theta$ the linear span formed by the vector

$$\underline{\varphi_k(\theta)}$$

is invariant under the action of M_k and K_k. This can be extended to the block-matrix $\mathcal{A}_k$ as follows: for all $\theta \in \Theta$ the linear span formed by the vectors

$$\begin{pmatrix} \underline{\varphi_k(\theta)} \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \underline{\varphi_k(\theta)} \end{pmatrix} \tag{5.2}$$

is invariant under the action of $\mathcal{A}_k$. Again, we can introduce the symbol:

$$\overline{\mathcal{A}_k}(\theta) = \begin{pmatrix} \overline{M_k}(\theta) & \overline{K_k}(\theta) \\ \overline{K_k}(\theta) & -\alpha^{-1}\overline{M_k}(\theta) \end{pmatrix}. \tag{5.3}$$

Here, the symbol is a 2-by-2 matrix and therefore it cannot be explained as an eigenvalue anymore. Note that the two vectors in (5.2) form a basis of a two-dimensional space. The symbol $\overline{\mathcal{A}_k}(\theta)$ is the representation of the block-matrix $\mathcal{A}_k$ with respect to that basis, i.e., we have the following relation.

Assume that $\overline{x_k}(\theta) = (\xi_1, \xi_2)$ is the representation of some $\underline{x}_k$ with respect to the basis (5.2), i.e., we have

$$\underline{x}_k = \xi_1 \begin{pmatrix} \underline{\varphi_k(\theta)} \\ 0 \end{pmatrix} + \xi_2 \begin{pmatrix} 0 \\ \underline{\varphi_k(\theta)} \end{pmatrix}.$$

Then the product $\mathcal{A}_k\, \underline{x}_k$ is also in the linear span spanned by the vectors given in (5.2). The representation of that product with respect to the bases is given by $\overline{\mathcal{A}_k}(\theta)\, \overline{x_k}(\theta)$, i.e., for $(\eta_1, \eta_2) = \overline{\mathcal{A}_k}(\theta)\, \overline{x_k}(\theta)$ we have

$$\mathcal{A}_k\, \underline{x}_k = \eta_1 \begin{pmatrix} \underline{\varphi_k(\theta)} \\ 0 \end{pmatrix} + \eta_2 \begin{pmatrix} 0 \\ \underline{\varphi_k(\theta)} \end{pmatrix}.$$

5.1.4 Symbol of the smoother

In the present subsection we determine the symbol of S_k, the iteration matrix of the smoother. As mentioned in the introduction, the analysis is presented for collective point smoothers (cf. Subsection 3.2.2) or, more precisely, for two smoothers of this class: the collective Jacobi smoother and the collective Gauss-Seidel smoother.

First, we discuss the collective Jacobi smoother. The preconditioner representing this smoother is given by $\hat{M}_k^{(jac)} := \operatorname{diag} M_k$ and $\hat{K}_k^{(jac)} := \operatorname{diag} K_k$. The preconditioners are diagonal matrices, therefore

$$\hat{M}_k^{(jac)}\ \underline{\varphi_k(\theta)} = \underbrace{\frac{2h_k}{3}}_{\overline{\hat{M}_k^{(jac)}}(\theta) :=}\ \underline{\varphi_k(\theta)} \quad \text{and} \quad \hat{K}_k^{(jac)}\ \underline{\varphi_k(\theta)} = \underbrace{\frac{2}{h_k}}_{\overline{\hat{K}_k^{(jac)}}(\theta) :=}\ \underline{\varphi_k(\theta)}$$

holds.

The preconditioner $\hat{\mathcal{A}}_k^{(jac)}$, representing the collective Jacobi smoother, is given by

$$\hat{\mathcal{A}}_k^{(jac)} = \begin{pmatrix} \hat{M}_k^{(jac)} & \hat{K}_k^{(jac)} \\ \hat{K}_k^{(jac)} & -\alpha^{-1}\hat{M}_k^{(jac)} \end{pmatrix}.$$

Analogous to Subsection 5.1.3, we can represent the symbol of $\hat{\mathcal{A}}_k^{(jac)}$ as 2-by-2 matrix with respect to the basis introduced in (5.2) and obtain

$$\overline{\hat{\mathcal{A}}_k^{(jac)}}(\theta) = \begin{pmatrix} \overline{\hat{M}_k^{(jac)}}(\theta) & \overline{\hat{K}_k^{(jac)}}(\theta) \\ \overline{\hat{K}_k^{(jac)}}(\theta) & -\alpha^{-1}\overline{\hat{M}_k^{(jac)}}(\theta) \end{pmatrix}.$$

As a consequence, we can also derive the symbol of the iteration matrix of the smoother:

$$\overline{S_k^{(jac)}}(\theta) = I - \tau\, \overline{\hat{\mathcal{A}}_k^{(jac)}}(\theta)^{-1}\overline{\mathcal{A}_k}(\theta).$$

A similar analysis can be worked out for the collective Gauss-Seidel smoother. Here, we assume that the nodes are updated in a consecutive way (from left to right). This smoother is represented by the preconditioners $\hat{M}_k^{(gs)}$ and $\hat{K}_k^{(gs)}$. These two matrices are the left-lower triangular part of the matrices M_k and K_k, respectively. The symbols are given by

$$\hat{M}_k^{(gs)}\ \underline{\varphi_k(\theta)} = \underbrace{\frac{(e^{-\mathrm{i}\theta}+4)h_k}{6}}_{\overline{\hat{M}_k^{(gs)}}(\theta) :=}\ \underline{\varphi_k(\theta)} \quad \text{and} \quad \hat{K}_k^{(gs)}\ \underline{\varphi_k(\theta)} = \underbrace{\frac{(-e^{-\mathrm{i}\theta}+2)}{h_k}}_{\overline{\hat{K}_k^{(gs)}}(\theta) :=}\ \underline{\varphi_k(\theta)}.$$

Analogously to the case of the collective Jacobi smoother, we have

$$\overline{\hat{\mathcal{A}}_k^{(gs)}}(\theta) = \begin{pmatrix} \overline{\hat{M}_k^{(gs)}}(\theta) & \overline{\hat{K}_k^{(gs)}}(\theta) \\ \overline{\hat{K}_k^{(gs)}}(\theta) & -\alpha^{-1}\overline{\hat{M}_k^{(gs)}}(\theta) \end{pmatrix}$$

and therefore, the symbol of the iteration matrix of the smoother reads as follows:

$$\overline{S_k^{(gs)}}(\theta) = I - \tau\, \overline{\hat{\mathcal{A}}_k^{(gs)}}(\theta)^{-1}\overline{\mathcal{A}_k}(\theta).$$

5.1.5 Symbol of the whole two-grid operator

As we are interested in the analysis of a whole two-grid step, we have also to take the coarse-grid correction into account. The coarse-grid correction operator consists of the restriction operator, the operator $\mathcal{A}_{k-1}$ on the coarser grid and the prolongation operator. First, we discuss the restriction operator I_k^{k-1}, which was defined in Chapter 3 as operator (matrix) acting on $\underline{x}_k = (\underline{y}_k, \underline{p}_k)$, i.e., on both variables. Certainly, we can restrict the state $\underline{y}_k$ and the adjoined state $\underline{p}_k$ separately, i.e., there is a restriction operator P_k^{k-1} such that $I_k^{k-1}(\underline{y}_k, \underline{p}_k) = (P_k^{k-1}\underline{y}_k, P_k^{k-1}\underline{p}_k)$ holds. The next step is the analysis of the operator P_k^{k-1}. One can verify that the restriction operator P_k^{k-1} maps the basis functions

$$\underline{\varphi_k(\theta)} \qquad \text{and} \qquad \underline{\varphi_k(\overline{\theta})} \tag{5.4}$$

to the same function

$$\underline{\varphi_{k-1}(2\theta)} \tag{5.5}$$

on the coarser grid for all $\theta \in \Theta^{(low)} := [\pi/2, \pi/2)$. Here and in what follows, $\overline{\theta}$ is given by

$$\overline{\theta} := \begin{cases} \theta + \pi & \text{for} \quad \theta < 0 \\ \theta - \pi & \text{for} \quad \theta \geq 0. \end{cases}$$

The same can be done for the prolongation operator P_{k-1}^k: this operator maps the function given in (5.5) to a linear combination of the functions in (5.4). Therefore, we cannot represent the two-grid correction operator with respect to the basis stated in (5.2) but with respect to the basis

$$\begin{pmatrix} \underline{\varphi_k(\theta)} \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \underline{\varphi_k(\theta)} \end{pmatrix}, \begin{pmatrix} \underline{\varphi_k(\overline{\theta})} \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ \underline{\varphi_k(\overline{\theta})} \end{pmatrix}, \tag{5.6}$$

see, e.g., Trottenberg [66] or Borzi, Kunisch and Kwak [12]. The symbols with respect to (5.6) of both, $\mathcal{A}_k$ and S_k, are block-diagonal:

$$\overline{\overline{\mathcal{A}_k}}(\theta) = \begin{pmatrix} \overline{\mathcal{A}_k}(\theta) & \\ & \overline{\mathcal{A}_k}(\overline{\theta}) \end{pmatrix} \in \mathbb{C}^{4\times 4}, \qquad \overline{\overline{S_k}}(\theta) = \begin{pmatrix} \overline{S_k}(\theta) & \\ & \overline{S_k}(\overline{\theta}) \end{pmatrix} \in \mathbb{C}^{4\times 4}. \tag{5.7}$$

The symbol of the intergrid transfer operator has a rectangular form, as the intergrid transfer operator maps from the basis given in (5.6) (for some θ) to the basis given in (5.2) (for 2θ), i.e., we have

$$\overline{I_{k-1}^k}(\theta) = \begin{pmatrix} \overline{P_{k-1}^k}(\theta) & 0 \\ 0 & \overline{P_{k-1}^k}(\theta) \\ \overline{P_{k-1}^k}(\bar{\theta}) & 0 \\ 0 & \overline{P_{k-1}^k}(\bar{\theta}) \end{pmatrix},$$

where $\overline{P_{k-1}^k}(\theta)$ is given by

$$\overline{P_{k-1}^k}(\theta) = \frac{1}{2}\left(\mathrm{e}^{-\theta \mathrm{i}} + 2 + \mathrm{e}^{\theta \mathrm{i}}\right).$$

Using these matrices, we can represent the symbol of the two-grid operator by

$$\overline{TG_k^{k-1}}(\theta) = \overline{\overline{S_k}}(\theta)^{\nu_{post}} \underbrace{\left(I - \overline{I_{k-1}^k}(\theta)\left(\overline{\mathcal{A}_{k-1}}(2\theta)\right)^{-1}\left(\overline{I_{k-1}^k}(\theta)\right)^T \overline{\overline{\mathcal{A}_k}}(\theta)\right)}_{\overline{CG_k^{k-1}}(\theta) =} \overline{\overline{S_k}}(\theta)^{\nu_{pre}}. \quad (5.8)$$

Here, the symbol $\overline{\mathcal{A}_{k-1}}(2\theta)$ is a 2-by-2 matrix, as introduced in (5.3). A similar analysis was done in Borzi, Kunisch and Kwak [12], cf. Theorem 5.1 in their work.

As mentioned earlier, we are interested in analyzing the norm of TG_k^{k-1}. The idea of local Fourier analysis is to compute the supremum of the norms of the symbols, i.e., we compute

$$\sup_{\theta \in \Theta} \left\| \overline{TG_k^{k-1}}(\theta) \right\|_{\overline{X}}, \quad (5.9)$$

where

$$\|A\|_{\overline{X}} := \left\| \begin{pmatrix} 1 & & & \\ & \alpha^{-1/2} & & \\ & & 1 & \\ & & & \alpha^{-1/2} \end{pmatrix} A \begin{pmatrix} 1 & & & \\ & \alpha^{1/2} & & \\ & & 1 & \\ & & & \alpha^{1/2} \end{pmatrix} \right\|_{\ell^2}.$$

This definition of the norm is motivated by the corresponding norm introduced in Theorem 12.

So, as mentioned above we have to take the supremum of (5.9) over all frequencies to obtain the convergence rate. Moreover, we are interested in an analysis that is robust in the grid size h_k and the choice of the parameter α. Therefore, we are interested in

$$q_{TG}(\tau) := \sup_{h_k > 0} \sup_{\alpha > 0} \sup_{\theta \in \Theta} \left\| \overline{TG_k^{k-1}}(\theta) \right\|_{\overline{X}}.$$

Here, the 4-by-4 matrix $\overline{TG_k^{k-1}}(\theta)$ and also its norm $\left\|\overline{TG_k^{k-1}}(\theta)\right\|_{\overline{X}}$ can be computed in a straight-forward way. The computation of the supremum is non-trivial but it can be done using tools from symbolic computation as outlined in the next section.

5.2 Quantifier elimination using cylindrical algebraic decomposition

In the present section we discuss how to compute the supremum of a given function using tools from symbolic computation. First we notice that the problem of computing the supremum can be equivalently rewritten as a general quantifier elimination problem. Let $D \subseteq \mathbb{R}^n$ and $f : \mathbb{R}^{m+n} \to \mathbb{R}$ be a function. Then, for fixed $y_1, \ldots, y_m$, the problem to compute

$$\sup_{(x_1,\ldots,x_n)\in D} f(x_1, \ldots, x_n, y_1, \ldots, y_m)$$

can be equivalently rewritten as follows: Find the smallest $\lambda \in \mathbb{R}$ such that

$$\forall\, (x_1, \ldots, x_n) \in D \;:\; f(x_1, \ldots, x_n, y_1, \ldots, y_m) \leq \lambda$$

is satisfied. This lower bound for λ can be derived easily if we are able to eliminate the quantifiers in this term. This can be done using quantifier elimination algorithms which allow to solve the following kind of problems (quantifier elimination problems).

Assume that a statement of the form

$$Q_1\, x_1 \ldots Q_n\, x_n : A(x_1, \ldots, x_n, y_1, \ldots, y_m)$$

is given, where the Q_i denote quantifiers (either $\forall$ or $\exists$) and $A(x_1, \ldots, x_n, y_1, \ldots, y_m)$ is a boolean combination of *polynomial* inequalities. The problem of finding an *equivalent, quantifier free formula* $B(y_1, \ldots, y_m)$ consisting of a boolean combination of polynomial inequalities depending only on the free variables is called a (polynomial) quantifier elimination problem. The first algorithm to solve this problem over the reals was given by Tarski [64] in the early 1950s. His method was practically not efficient. G. Collins' cylindrical algebraic decomposition, cf. Collins [27], makes it possible to carry out non-trivial computations in a reasonable amount of time. Modern implementations of this approach were developed by Brown [25], Seidel and Sturm [56], Strzeboński [59] and others.

For illustrating the technique of cylindrical algebraic decomposition, we consider the following simple example: Determine for $z > 0$

$$g(z) := \sup_{0<x<1} \sup_{0<y<1} \frac{x}{y+z} + \frac{y}{x+z}. \tag{5.10}$$

This can be rewritten as a quantified formula by introducing an additional variable λ:

$$z > 0 \wedge \forall\, 0 < x < 1 \,\forall\, 0 < y < 1 : \frac{x}{y+z} + \frac{y}{x+z} \leq \lambda,$$

or equivalently,

$$\forall\, x \,\forall\, y : z > 0 \wedge \left[0 < x < 1 \wedge 0 < y < 1 \Rightarrow \frac{x}{y+z} + \frac{y}{x+z} \leq \lambda\right].$$

This is equivalent to

$$\forall\, x \,\forall\, y : z > 0 \wedge [0 < x < 1 \wedge 0 < y < 1 \Rightarrow x(x+z) + y(y+z) \leq \lambda(x+z)(y+z)]\,.$$

Here we use that $x > 0$, $y > 0$ and $z > 0$. (Such a rewriting in polynomial form is also possible if the denominator cannot be guaranteed to be positive. However, in this case the formula may be more complicated).

Here we have $A(x, y, z, \lambda) \equiv z > 0 \wedge [0 < x < 1 \wedge 0 < y < 1 \Rightarrow x(x+z) + y(y+z) \leq \lambda(x+z)(y+z)]$. We use Mathematica's command Resolve to perform a cylindrical algebraic decomposition:

In[1]= **Resolve[ForAll[x, 0 < x < 1, ForAll[y, 0 < y < 1, z > 0&&$x(x+z) + y(y+z) \leq \lambda(x+z)(y+z)$]], {$z$, λ}, Reals]**

Out[1]= $\left(0 < z \leq 1 \&\& \lambda \geq \frac{1}{z}\right) \| \left(z > 1 \&\& \lambda \geq \frac{2}{1+z}\right)$

This means, we obtain $B(z, \lambda) \equiv \left(0 < z \leq 1 \wedge \lambda \geq \frac{1}{z}\right) \vee \left(z > 1 \wedge \lambda \geq \frac{2}{1+z}\right)$. We see that in the result only the free variables λ and z (which were not fixed by the quantifiers) appear. The bound variables x and y (variables that are fixed by the quantifiers) do not appear anymore. Consequently, in cases where no free variables appear in the input, the result is one of the logical constants True or False.

When executing the algorithm first the quantifier free part of the formula is considered, i.e., in the example above the inequalities $z > 0$, $0 < x < 1$, $0 < y < 1$ and $x(x+z) + y(y+z) \leq \lambda(x+z)(y+z)$. The given polynomials define a natural decomposition of the real space (in the example $\mathbb{R}^4$) into maximal connected cells on which the polynomials are sign invariant. This decomposition is then further refined by the algorithm to obtain cells on which the polynomials are not only sign invariant, but the cells are also *cylindric*, i.e., every cell $C \subseteq \mathbb{R}^n$ has the following form:

- $C = \{(x, y) \in \mathbb{R}^{n-1} \times \mathbb{R} \,:\, x \in D \text{ and } \underline{\psi}(x) < y < \overline{\psi}(x)\}$ or
- $C = \{(x, y) \in \mathbb{R}^{n-1} \times \mathbb{R} \,:\, x \in D \text{ and } y = \psi(x)\}$,

where $\underline{\psi}$, $\overline{\psi}$ and $\psi : \mathbb{R}^{n-1} \to \mathbb{R} \cup \{-\infty, \infty\}$ and D is a *cylindric* cell in $\mathbb{R}^{n-1}$. On $\mathbb{R}$, cylindric cells are open intervals or single points.

The requirement that all cells are cylindrical assumes the variables to be ordered. This ordering is fixed for the bound variables by the order of the quantifiers and by the user (or the implementation) for the free variables. In this sense one may consider variables as being on the bottom (or innermost) level or on higher levels of the resulting CAD. Once such a cylindrical decomposition is obtained the quantifiers can be eliminated by considering each of the cells in an order determined by the quantifiers. The result is a formula where all the bound variables have been eliminated, and the description of the cells where the formula holds is given solely in terms of the free variables as shown in the example above.

This procedure may be very costly depending heavily on the input parameters such as the number of polynomial inequalities, the polynomial degrees and the number of variables. In the worst case it is doubly exponential in the number of variables and this worst case bound is not only met in theory, but often experienced in practice. As we will see below, already for the one dimensional analysis suitable substitutions of the variables are applied in order to speed up the computations. These substitutions aim at reducing the number of variables on the one hand and lowering the polynomial degrees on the other hand. Although it might seem a high price to pay, the gain is an *optimal* bound for the given formula that is determined by a *proving* procedure that is not approximate in any way.

For the forthcoming analysis of the two (or even three) dimensional case, further simplifications will be necessary because of the increase in both, the number of unknowns as well as the polynomial degrees of the given formulas.

Return to the problem of computing $g(z)$. As mentioned, the supremum is the smallest upper bound, i.e., the smallest λ that satisfies

$$B(z,\lambda) \equiv \left(0 < z \leq 1 \wedge \lambda \geq \frac{1}{z}\right) \vee \left(z > 1 \wedge \lambda \geq \frac{2}{1+z}\right). \tag{5.11}$$

Therefore, $g(z)$ is a piecewise linear function, given by

$$g(z) = \begin{cases} \frac{1}{z} & \text{for } 0 < z \leq 1 \\ \frac{2}{1+z} & \text{for } z > 0 \end{cases}.$$

Note that for the interpretation of as piecewise defined function, the prescribed ordering of the variables is of importance.

In this section, we have seen that we are able to resolve the supremum of a rational function. In the further sections of the chapter, we will apply this strategy like a black-box to resolve suprema of our interest.

5.3 An analysis based on smoothing rates

The goal of this section is the computation of sharp upper bounds for the convergence rate of the two-grid method. The first step is the computation of the smoothing rate because the computation of the smoothing rates is rather simple and is a good starting point for doing the CAD computations presented in the last section. The smoothing rate has been introduced as smoothing factor in equation (3.8) in Brandt [19]. Also in Borzi, Kunisch and Kwak [12] the smoothing rates were derived. Contrary to the present work, they approximated these rates numerically.

Certainly, the smoothing rates we will compute do not give any information about the convergence of the overall multigrid method (or two-grid method) directly. We will present a possibility to construct – based on the information on the smoothing rate – an upper bound for the convergence of the two-grid method.

Often (cf. Borzi, Kunisch and Kwak [12] or Trottenberg [66]) such an analysis is not done and the two-grid convergence rates are computed directly (without using the computed smoothing rates). Since the whole two-grid operator is considered, we call this an all-at-once analysis. We will give such an analysis in Section 5.4. The big advantage of the all-at-once analysis is the fact that sharp upper bounds of the convergence rate of the two-grid method is computed whereas the analysis presented in this section gives relatively rough upper bounds for the convergence rate.

Nonetheless, the separation of the analysis for the smoother and the coarse-grid correction done in this section simplifies the comparison of different kinds of smoothers and the analysis for varying numbers of smoothing steps ν. Moreover, the separation decomposes the original problem to smaller subproblems which seems to be the key for extending the results presented here to higher dimensions. (This is not an issue if convergence rates are approximated numerically because there the complexity of the analysis is not growing exponentially with the size of the problem.)

5.3.1 A rigorous justification for the use of smoothing rates

As already mentioned, the concept of smoothing rates, that is used in Subsection 5.3.2, is well-known. The following theorem states that the combination of the smoothing rate

q_{SM} and a complementary result on the coarse grid correction directly yields an upper bound for the convergence rate. The proof of this theorem shows that the concept of the smoothing property follows naturally from a splitting of the two-grid operator.

Theorem 67 *Consider the two-grid method, introduced in Section 5.1, with $\nu/2$ pre- and $\nu/2$ post-smoothing steps. Assume that the smoother satisfies*

$$\sup_{\theta\in\Theta^{(low)}} \left\|\overline{S_k}(\theta)\right\|_{\overline{X}} \leq 1. \tag{5.12}$$

Then, the convergence rate

$$q_{TG} := \sup_{\theta\in\Theta^{(low)}} \left\|\overline{TG_k^{k-1}}(\theta)\right\|_{\overline{X}}$$

can be bounded as follows

$$q_{TG} \leq \widetilde{q_{TG}}(q_{SM}^{\nu/2}),$$

where

$$\widetilde{q_{TG}}(q) := \sup_{\theta\in\Theta^{(low)}} \left\|\mathcal{I}(q)\overline{CG_k^{k-1}}(\theta)\mathcal{I}(q)\right\|_{\overline{X}}, \tag{5.13}$$

$$\mathcal{I}(q) := \begin{pmatrix} I & \\ & qI \end{pmatrix}, \tag{5.14}$$

$$q_{SM} := \sup_{\theta\in\Theta^{(high)}} \left\|\overline{S_k}(\theta)\right\|_{\overline{X}}, \tag{5.15}$$

$\Theta^{(low)} := [-\pi/2, \pi/2)$ *and* $\Theta^{(high)} := [-\pi, \pi)\backslash[-\pi/2, \pi/2)$.

Proof: Using (5.8), the semi-multiplicativity of operator norms, (5.7) and (5.14), we obtain

$$\begin{aligned}
\left\|\overline{TG_k^{k-1}}(\theta)\right\|_{\overline{X}} &= \left\|\overline{\overline{S_k}}(\theta)^{\nu/2}\,\overline{CG_k^{k-1}}(\theta)\,\overline{\overline{S_k}}(\theta)^{\nu/2}\right\|_{\overline{X}} \\
&= \left\|\overline{\overline{S_k}}(\theta)^{\nu/2}\,\mathcal{I}(q^{-\nu/2})\,\mathcal{I}(q^{\nu/2})\,\overline{CG_k^{k-1}}(\theta)\,\mathcal{I}(q^{\nu/2})\,\mathcal{I}(q^{-\nu/2})\,\overline{\overline{S_k}}(\theta)^{\nu/2}\right\|_{\overline{X}} \\
&\leq \left\|\mathcal{I}(q^{\nu/2})\,\overline{CG_k^{k-1}}(\theta)\,\mathcal{I}(q^{\nu/2})\right\|_{\overline{X}} \left\|\mathcal{I}(q^{-\nu/2})\,\overline{\overline{S_k}}(\theta)^{\nu/2}\right\|_{\overline{X}}^2 \\
&= \left\|\mathcal{I}(q^{\nu/2})\,\overline{CG_k^{k-1}}(\theta)\,\mathcal{I}(q^{\nu/2})\right\|_{\overline{X}} \left\|\mathcal{I}(q)^{-\nu/2} \begin{pmatrix} \overline{S_k}(\theta)^{\nu/2} & \\ & \overline{S_k}(\overline{\theta})^{\nu/2} \end{pmatrix}\right\|_{\overline{X}}^2 \\
&= \left\|\mathcal{I}(q^{\nu/2})\,\overline{CG_k^{k-1}}(\theta)\,\mathcal{I}(q^{\nu/2})\right\|_{\overline{X}} \max\left\{\left\|\overline{S_k}(\theta)^{\nu/2}\right\|_{\overline{X}}, q^{-\nu/2}\left\|\overline{S_k}(\overline{\theta})^{\nu/2}\right\|_{\overline{X}}\right\}^2 \\
&\leq \left\|\mathcal{I}(q^{\nu/2})\,\overline{CG_k^{k-1}}(\theta)\,\mathcal{I}(q^{\nu/2})\right\|_{\overline{X}} \max\left\{\left\|\overline{S_k}(\theta)\right\|_{\overline{X}}, q^{-1}\left\|\overline{S_k}(\overline{\theta})\right\|_{\overline{X}}\right\}^{\nu}.
\end{aligned} \tag{5.16}$$

If we choose

$$q := q_{SM} = \sup_{\theta\in\Theta^{(high)}} \left\|\overline{S_k}(\theta)\right\|_{\overline{X}},$$

we obtain

$$q^{-1} \left\| \overline{S_k}(\overline{\theta}) \right\|_{\overline{X}} \leq 1$$

for all $\theta \in \Theta^{(low)}$ because $\theta \in \Theta^{(low)}$ implies $\overline{\theta} \in \Theta^{(high)}$. Since we have also (5.12), we obtain

$$\max \left\{ \left\| \overline{S_k}(\theta) \right\|_{\overline{X}}, q^{-1} \left\| \overline{S_k}(\overline{\theta}) \right\|_{\overline{X}} \right\}^{\nu} \leq 1$$

and therefore due to (5.16) finally

$$q_{TG} = \left\| \overline{TG_k^{k-1}}(\theta) \right\|_{\overline{X}} \leq \widetilde{q_{TG}}(q_{SM}^{\nu/2}) = \left\| \mathcal{I}(q_{SM}^{\nu/2}) \, \overline{CG_k^{k-1}}(\theta) \, \mathcal{I}(q_{SM}^{\nu/2}) \right\|_{\overline{X}}$$

for all $\theta \in \Theta^{(low)}$. □

We will see in Subsection 5.3.2 that (5.12) holds and the smoothing rate $q_{SM} < 1$ can be computed. In Subsection 5.3.3, we will see that, provided q_{SM} small enough, an optimal and robust convergence result can be shown, i.e., that the convergence rate is bounded by $\widetilde{q_{TG}}(q_{SM}^{\nu/2}) < 1$, which is independent in h_k and α.

5.3.2 Smoothing rates

First, we analyze the collective Jacobi smoother. We compute

$$q_{SM}(\tau) := \sup_{\theta \in \Theta^{(high)}} \sup_{h_k > 0} \sup_{\alpha > 0} \sigma(\theta, h_k, \alpha, \tau), \tag{5.17}$$

where

$$\sigma(\theta, h_k, \alpha, \tau) := \Big\| \underbrace{I - \tau \left(\overline{\hat{\mathcal{A}}_k^{(jac)}}(\theta) \right)^{-1} \overline{\mathcal{A}_k}(\theta)}_{\overline{S}_k^{(jac)}(\theta) :=} \Big\|_{\overline{X}}.$$

The computation of $\sigma(\theta, h_k, \alpha, \tau)$ is straight forward. We obtain

$$\sigma^2(\theta, h_k, \alpha, \tau) = \frac{h_k^4((\cos\theta + 2)\tau - 2)^2 + 36\alpha((\cos\theta - 1)\tau + 1)^2}{4\left(h_k^4 + 9\alpha\right)}.$$

Here, due to the presence of a trigonometric function $\cos\theta$, the function σ^2 is not a rational function. The cosine is eliminated by replacing $\cos\theta$ by some variable c. As $\theta \in \Theta^{(high)}$ is equivalent to $c := \cos\theta \in [-1, 0]$, we obtain

$$q_{SM}^2(\tau) = \sup_{c \in [-1,0]} \sup_{h_k > 0} \sup_{\alpha > 0} \widetilde{\sigma}^2(c, h_k, \alpha, \tau), \tag{5.18}$$

where

$$\widetilde{\sigma}^2(c, h_k, \alpha, \tau) := \frac{h_k^4((c + 2)\tau - 2)^2 + 36\alpha((c - 1)\tau + 1)^2}{4\left(h_k^4 + 9\alpha\right)}.$$

Resolving (5.18) is done as outlined in Section 5.2. So we have to find the smallest λ that satisfies

$$\forall c \in [-1,0] \ : \ \forall h_k > 0 \ : \ \forall \alpha > 0 \ \widetilde{\sigma}^2(c, h_k, \alpha, \tau) \leq \lambda.$$

Here, the quantifies can be eliminated with Mathematica's Resolve command in less than a second. We obtain the following equivalent formula:

$$\begin{aligned}\left(\tau \leq 0 \wedge 4\tau^2 - 4\tau + 1 \leq \lambda\right) \vee \left(0 < \tau \leq \frac{4}{5} \wedge \frac{1}{4}\left(\tau^2 - 4\tau + 4\right) \leq \lambda\right) \\ \vee \left(\frac{4}{5} < \tau \wedge 4\tau^2 - 4\tau \leq \lambda\right).\end{aligned}$$

Therefore, we obtain

$$q_{SM}^2(\tau) = \begin{cases} 4\tau^2 - 4\tau + 1 & \text{for} \quad \tau \leq 0 \\ \frac{1}{4}\left(\tau^2 - 4\tau + 4\right) & \text{for} \quad 0 < \tau \leq \frac{4}{5} \\ 4\tau^2 - 4\tau + 1 & \text{for} \quad \frac{4}{5} < \tau \end{cases}. \tag{5.19}$$

If we take the square root of (5.19) and restrict ourselves to the relevant range $\tau \in [0,1]$, we obtain the following result.

Theorem 68 *The smoothing rate for the collective Jacobi smoother is given by*

$$q_{SM}(\tau) = \begin{cases} \frac{1}{2}(2-\tau) & \text{for} \quad 0 \leq \tau \leq \frac{4}{5} \\ 2\tau - 1 & \text{for} \quad \frac{4}{5} < \tau \leq 1 \end{cases} \tag{5.20}$$

for all $\tau \in [0,1]$.

The graph of the function q_{SM} can be seen in Figure 5.1. $q_{SM}(\tau)$ takes its minimum for $\tau = \frac{4}{5}$ with value $q_{SM}\left(\frac{4}{5}\right) = \frac{3}{5}$. For the choice $\tau = \frac{1}{2}$, we obtain $q_{SM}\left(\frac{1}{2}\right) = \frac{3}{4}$.

A smoothing analysis in a similar setting has been carried out in Borzi, Kunisch and Kwak [12], where the authors obtain estimates for smoothing and convergence rates using numerical interpolation. To the knowledge of the author (5.20) provide the first rigorously proven sharp bounds for the smoothing rate (cf. Pillwein and Takacs [48]).

We have to show also (5.12) for all h_k and α, i.e., we have to show

$$\sup_{\theta \in \Theta^{(low)}} \sup_{h_k > 0} \sup_{\alpha > 0} \sigma(\theta, h_k, \alpha, \tau) \leq 1, \tag{5.21}$$

which can be done as above. Here, we obtain

$$\sup_{h_k > 0} \sup_{\alpha > 0} \sigma(\theta, h_k, \alpha, \tau) = \begin{cases} 1 - \frac{3\tau}{2} & \text{for} \quad \tau < 0 \\ 1 & \text{for} \quad \tau \geq 0 \end{cases},$$

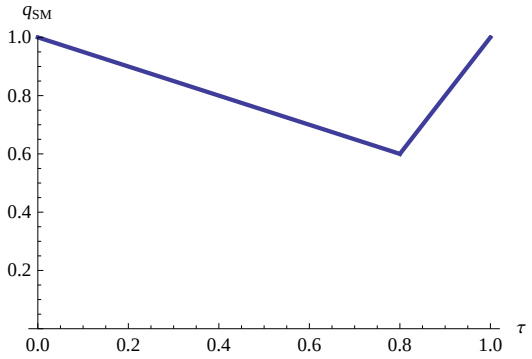


Figure 5.1: Smoothing factor depending on damping parameter τ

i.e., (5.21) holds for all $\tau \in [0,1]$.

We can develop the smoothing analysis for the collective Gauss-Seidel smoother similar to the analysis above. The damping parameter is fixed: $\tau = 1$. Therefore, the smoothing rate does not depend on τ anymore:

$$q_{SM} := \sup_{\theta \in \Theta^{(high)}} \sup_{h_k > 0} \sup_{\alpha > 0} \sigma(\theta, h_k, \alpha),$$

where

$$\sigma(\theta, h, \alpha) := \left\| \overline{S_k}^{(gs)}(\theta) \right\|_{\overline{X}}.$$

Here, σ can be computed completely analogous to the previous case:

$$\sigma^2(\theta, h_k, \alpha) := \frac{\left(h_k^4 + 36\alpha\right)\left((17+8c)h_k^4 + 72h_k^2\alpha^{1/2}|\sin\theta| + 36(5-4\cos\theta)\alpha\right)}{(17+8\cos\theta)^2 h_k^8 + 72(40\cos^2\theta - 28\cos\theta + 13)h_k^4\alpha + 1296(5-4\cos\theta)^2\alpha^2}.$$

In this formula, the occurrences of $\cos\theta$ and $\sin\theta$ are replaced by c and s, respectively. For an equivalent rewriting, we have to require Pythagoras' identity $s^2+c^2 = 1$ explicitly as constraint. The fact that $\theta \in \Theta^{(low)}$ is equivalent to $c \in [-1, 0]$ (as in the case of collective Jacobi relaxation). The absolute value is eliminated as follows. Because σ does not depend on the sign of $s = \sin\theta$, we can restrict ourselves to assuming $s \geq 0$, which allows to replace $|s|$ by s. Moreover, we replace $\alpha^{1/2}$ by $\widetilde{\alpha} > 0$. Using these rewritings, the final formula for q_{SM} reads as follows.

$$q_{SM}^2 = \sup_{(s,c)\in D} \sup_{h_k > 0} \sup_{\widetilde{\alpha} > 0} \frac{\left(h_k^4 + 36\widetilde{\alpha}^2\right)\left((17+8c)h_k^4 + 72h_k^2\widetilde{\alpha}s + 36(5-4c)\widetilde{\alpha}^2\right)}{(17+8c)^2 h_k^8 + 72(40c^2 - 28c + 13)h_k^4\widetilde{\alpha}^2 + 1296(5-4c)^2\widetilde{\alpha}^4},$$

where $D := \{(s,c) \in \mathbb{R}^2 \ : \ s^2 + c^2 = 1,\ c \leq 0,\ s \geq 0\}$.

We solve the problem using Mathematica's Resolve and obtain after about twenty minutes a quantifier free formula. We obtain the following statement.

Theorem 69 *The smoothing rate for the collective Gauss-Seidel smoother is given by*

$$q_{SM} = \tfrac{1}{7}\left(3+\sqrt{2}\right) \approx 0.63.$$

Even though twenty minutes are not a very long time to wait for a result that needs to be obtained only once, it still seems too long for such a simple formula. We can speed up the calculation significantly by reducing both, the number of variables and the degrees of the polynomials, by substituting $\widetilde{\alpha}/h_k^2 = \alpha^{1/2}/h_k^2$ by a new variable $\eta := \alpha^{1/2}/h_k^2$. This substitution reduces the formula for q_{SM} to

$$q_{SM}^2 = \sup_{(s,c)\in\widetilde{D}} \sup_{\eta>0} \frac{\left(1+36\eta^2\right)\left((17+8c)+72\eta s+36(5-4c)\eta^2\right)}{(17+8c)^2+72(40c^2-28c+13)\eta^2+1296(5-4c)^2\eta^4}.$$

Based on this representation Mathematica's Resolve command is able to derive q_{SM} within about twenty seconds.

We have to show also (5.12) for all h_k and α, i.e., we have to show

$$\sup_{\theta\in\Theta^{(low)}} \sup_{h_k>0} \sup_{\alpha>0} \sigma(\theta, h_k, \alpha, \tau) \leq 1.$$

We can compute the supremum using the rewritings introduced above. Mathematica's Resolve command terminates within some seconds. We obtain that the supremum is equal to 1.

An alternative approach for showing (5.12) is just to verify

$$\forall\theta\in\Theta^{(low)} \ : \ \forall h_k>0 \ : \ \forall\alpha>0 \ : \ \sigma(\theta,h_k,\alpha,\tau)\leq 1.$$

Mathematica's quantifier elimination algorithm can be applied directly to such a problem (after applying the rewritings introduced above) and yields the logical constant True, which shows that the supremum in (5.12) is smaller or equal 1.

5.3.3 Two-grid convergence rate

In this subsection we follow the approach introduced in Theorem 67 and derive $\widetilde{q_{TG}}$, introduced in (5.13). Since we take also the supremum with respect to h_k and α, the formula for $\widetilde{q_{TG}}$ reads as follows

$$\widetilde{q_{TG}}(q) := \sup_{\theta\in\Theta^{(low)}} \sup_{h_k>0} \sup_{\alpha>0} \left\| \mathcal{I}(q)\,\overline{CG_k^{k-1}}(\theta)\,\mathcal{I}(q) \right\|_{\overline{X}}.$$

We compute $\sigma(\theta, h_k, \alpha, q) := \left\| \mathcal{I}(q)\, \overline{CG_k^{k-1}}(\theta)\, \mathcal{I}(q) \right\|_{\overline{X}}$ first. We have

$$\begin{aligned}
&\left\| \mathcal{I}(q)\, \overline{CG_k^{k-1}}(\theta) \mathcal{I}(q) \right\|_{\overline{X}}^2 \\
&\qquad = \lambda_{\max}\Big(\underbrace{\mathcal{L}_k^{-1/2} \mathcal{I}(q)\, \overline{CG_k^{k-1}}(\theta) \mathcal{I}(q)\, \mathcal{L}_k \mathcal{I}(q)\, \overline{CG_k^{k-1}}(\theta) \mathcal{I}(q)\, \mathcal{L}_k^{-1/2}}_{\mathcal{N}_k :=} \Big).
\end{aligned}$$

The matrix $\mathcal{N}_k$ can be computed in a straight-forward way and we obtain

$$\mathcal{N}_k = \frac{P(h_k, \alpha, c, q)}{Q(h_k, \alpha, c, q)} \begin{pmatrix} 1 & 0 & \frac{(c+1)q}{c-1} & 0 \\ 0 & 1 & 0 & \frac{(c+1)q}{c-1} \\ \frac{(c+1)q}{c-1} & 0 & \left(\frac{(c+1)q}{c-1}\right)^2 & 0 \\ 0 & \frac{(c+1)q}{c-1} & 0 & \left(\frac{(c+1)q}{c-1}\right)^2 \end{pmatrix},$$

where

$$\begin{aligned}
P(h_k, \alpha, c, q) &:= (h_k^4((c-2)^2(c-1)^4 + (c-1)^2(c+1)^2(c+2)^2 q \\
&\qquad + 36(c-1)^4(c+1)^2(1+q)\alpha), \\
Q(h_k, \alpha, c, q) &:= 16((2c^2+1)^2 h_k^4 + 9(c^2-1)^2\alpha).
\end{aligned}$$

We can compute the spectral radius of $\mathcal{N}_k$ and obtain

$$\sigma^2(h_k, \alpha, c, q) = \left(1 + \left(\frac{(1+c)q}{c-1}\right)^2\right) \frac{P(h_k, \alpha, c, q)}{Q(h_k, \alpha, c, q)}.$$

We are interested in computing

$$\widetilde{q_{TG}}^2(q) = \sup_{h_k>0} \sup_{\alpha>0} \sup_{0\le c\le 1} \sigma^2(h_k, \alpha, c, q).$$

In principle, this can be resolved using CAD. Unfortunately, the computation does not terminate within a reasonable time. We can simplify the problem here in a similar way as it was done for the smoothing property: Also in this subsection, the function σ does not depend on h_k and α individually but only on $\eta := \alpha/h_k^4$. Therefore, we could try to compute

$$\widetilde{q_{TG}}^2(q) = \sup_{\eta>0} \sup_{0\le c\le 1} \sigma^2(1, \eta, c, q).$$

using CAD. But also this is not possible within a reasonable time. Therefore, we have to use further information. Observe that

$$\sigma^2(1, \eta, c, q) = \frac{A_1(c,q) + \eta A_2(c,q)}{B_1(c,q) + \eta B_2(c,q)},$$

where

$$
\begin{aligned}
A_1(c,q) &:= ((c-2)^2(c-1)^2 + (c+1)^2(c+2)^2 q))((c-1)^2 + (c+1)^2 q) \\
A_2(c,q) &:= 36(c-1)^2(c+1)^2(q+1)((c-1)^2 + (c+1)^2 q) \\
B_1(c,q) &:= 16(2c^2+1)^2 \\
B_2(c,q) &:= 144(c^2-1)^2.
\end{aligned}
$$

It is easy to see that

$$
\sup_{\eta>0} \frac{A_1(c,q) + \eta A_2(c,q)}{B_1(c,q) + \eta B_2(c,q)} = \max\left\{ \frac{A_1(c,q)}{B_1(c,q)}, \frac{A_2(c,q)}{B_2(c,q)} \right\}
$$

holds in general, i.e., we take the maximum over the cases $\eta = 0$ and $\eta \to \infty$. Now we are able to compute $\sup_{0\le c\le 1} \frac{A_1(c,q)}{B_1(c,q)}$ and $\sup_{0\le c\le 1} \frac{A_2(c,q)}{B_2(c,q)}$ within some minutes and obtain the following result.

Theorem 70 *The convergence rate for the two-grid method is given by*

$$
\widetilde{q_{TG}}(q) = \begin{cases} \frac{1+q}{2} & \text{for } 0 \le q \le \frac{1}{3} \\ \sqrt{q(1+q)} & \text{for } \frac{1}{3} < q \end{cases}
$$

for all $q > 0$.

The function $\widetilde{q_{TG}}(q)$ is visualized in Figure 5.2. Since $\widetilde{q_{TG}}(0) = \frac{1}{2}$, using this method a convergence rate better that $\frac{1}{2}$ cannot be shown. To obtain convergence at all, we have to guarantee that $q := q_{SM}^{\nu} < \frac{\sqrt{5}-1}{2} \approx 0.62$ holds.

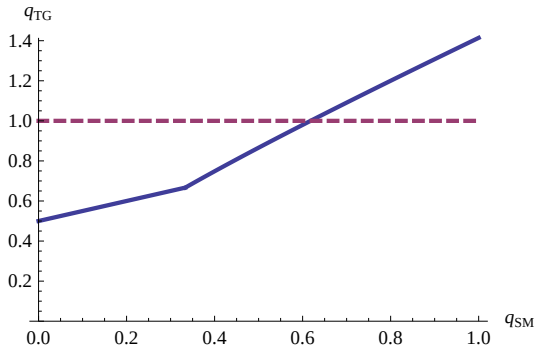


Figure 5.2: Convergence rate depending on smoothing rate

We can combine the results of this subsection and the last subsection to obtain upper bounds for the convergence rate of the two-grid method, i.e., we obtain

$$q_{TG} \leq \widetilde{q_{TG}}(q_{SM}^{\nu}).$$

In case of collective Jacobi iteration, we obtain the following result.

Corollary 71 *If $\nu = \nu_{pre} + \nu_{post} = 2 + 2$ smoothing steps are applied and collective Jacobi iteration is chosen as smoother, the convergence rate can be bounded as follows.*

$$q_{TG}(\tau) \leq \widetilde{q_{TG}}(q_{SM}^2(\tau)) = \begin{cases} \frac{1}{4}(2-\tau)\sqrt{8+(\tau-4)\tau} & \text{for } 0 \leq \tau \leq \frac{4}{5} \\ (2\tau-1)\sqrt{2+4(\tau-1)\tau} & \text{for } \frac{4}{5} < \tau \leq 1 \end{cases}.$$

The bound of the two-grid convergence rate, $\widetilde{q_{TG}}(q_{SM}^2(\tau))$, is visualized in Figure 5.3. Here, the optimal choice for τ follows directly from the smoothing analysis (as $\widetilde{q_{TG}}$ is just a monotone increasing function). The optimal choice is $\tau = \frac{4}{5}$ which leads to $\widetilde{q_{TG}}\left(q_{SM}^2\left(\frac{4}{5}\right)\right) = \frac{1}{25}\left(3\sqrt{34}\right) \approx 0.70$. For the choice $\tau = \frac{1}{2}$, we obtain $\widetilde{q_{TG}}\left(q_{SM}^2\left(\frac{1}{2}\right)\right) = \frac{15}{16} \approx 0.94$.

In case of collective Gauss-Seidel iteration, we obtain the following result.

Corollary 72 *If $\nu = \nu_{pre} + \nu_{post} = 2 + 2$ smoothing steps are applied and collective Gauss-Seidel iteration is chosen as smoother, the convergence rate can be bounded as follows.*

$$q_{TG} \leq \widetilde{q_{TG}}(q_{SM}^2) = \frac{1}{49}\sqrt{732 + 426\sqrt{2}} \approx 0.75.$$

Note that for both, the collective Jacobi smoother and the collective Gauss-Seidel smoother, we are not able to show convergence using the approach proposed in this section if only $\nu = \nu_{pre} + \nu_{post} = 1 + 1$ smoothing steps are applied. This case is covered by the analysis we will present in the next section.

5.4 An all-at-once analysis

As mentioned in the beginning of the last section, we can also analyze the complete two-grid operator in one step, i.e., we can compute

$$q_{TG}^2(\tau) = \sup_{h_k>0} \sup_{\alpha>0} \sup_{\theta\in\Theta} \left\| \overline{TG_k^{k-1}}(\theta) \right\|_{\overline{X}}^2 \tag{5.22}$$

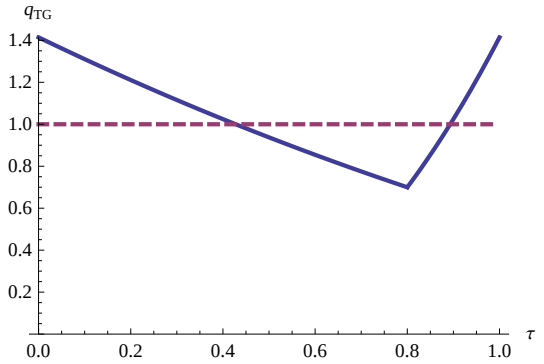


Figure 5.3: Convergence rate depending on damping parameter τ for 2 pre- and 2 post-smoothing steps

directly for some fixed smoother and some fixed ν. Note that here, contrary to the last section, not only an upper bound for the overall convergence rate is computed but a sharp upper bound is determined.

The term $\left\|\overline{TG_k^{k-1}}(\theta)\right\|_{\overline{X}}$ depends on all of the variables, τ, h_k, α and θ. The norm can be computed in a straight-forward way. We obtain

$$\left\|\overline{TG_k^{k-1}}(\theta)\right\|_{\overline{X}}^2 = \sigma\left(\tau, \frac{h_k^4}{\alpha}, \cos^2(\theta)\right),$$

where

$$\sigma(\tau,\eta,\gamma) = \frac{P_1(\tau,\eta,\gamma)P_2(\tau,\eta,\gamma)}{64(9+\eta)^2(9(\gamma-1)^2+\eta(1+2\gamma)^2)},$$

with

$$\begin{aligned}
P_1(\tau,\eta,\gamma) &:= \eta\left(\gamma^2\tau^2+\gamma\left(4-3\tau^2\right)+4(\tau-1)^2\right)\\
&\quad + 36\left(\gamma^2\tau^2+\gamma\left(6\tau^2-6\tau+1\right)+(\tau-1)^2\right)\\
P_2(\tau,\eta,\gamma) &:= \eta^2\left(\gamma^3\tau^2+\gamma^2\left(4+16\tau-7\tau^2\right)+\gamma\left(8\tau^2-56\tau+52\right)+16(\tau-1)^2\right)\\
&\quad + 36\eta\left(2\gamma^3\tau^2+\gamma^2\left(28\tau^2-22\tau+5\right)+\gamma\left(34\tau^2-34\tau+5\right)+8(\tau-1)^2\right)\\
&\quad + 1296(\gamma-1)^2\left((\gamma+1)\tau^2-2\tau+1\right).
\end{aligned}$$

Consequently, the equation (5.22) can be rewritten using the function σ and we obtain

$$q_{TG}^2(\tau) = \sup_{\eta>0}\sup_{0\le\gamma\le1}\sigma(\tau,\eta,\gamma). \tag{5.23}$$

Again, we have to resolve the supremum of a rational function which can be done as outlined in Section 5.2. Unfortunately, a direct application of Mathematica's Resolve command to the quantified formula representing the problem (5.23) does not terminate within a reasonable time.

Therefore, the problem has to be simplified further. Here, contrary to the case of the last section, numerator and denominator are not linear in η. Therefore, we cannot apply the strategy used there, directly. We use a slightly different approach: we guess the convergence rate using the samples $\eta := 0$ and $\eta \to \infty$ as a first step and show that the guess is correct as a second step.

As outlined, by sampling we obtain

$$\begin{aligned}
\sigma_0(\tau,\gamma) &:= \sigma(\tau,0,\gamma) = (1+\tau(-2+\tau+\gamma\tau))((\tau-1)^2+\gamma^2\tau^2+\gamma(1+6(\tau-1)\tau)),\\
\sigma_\infty(\tau,\gamma) &:= \lim_{\eta\to\infty}\sigma(\tau,\eta,\gamma)\\
&= \frac{1}{64(1+2\gamma)^2}(4(\tau-1)^2+\gamma^2\tau^2+\gamma(4-3\tau^2))(16(\tau-1)^2+\gamma^3\tau^2\\
&\quad +\gamma^2(4+16\tau-7\tau^2)+\gamma(52-56\tau+8\tau^2)).
\end{aligned}$$

We compute the supremum using Mathematica's Resolve command for both cases separately and obtain

$$q_0^2(\tau) := \sup_{0\le\gamma\le 1}\sigma_0(\tau,\gamma) \qquad \text{and} \qquad q_\infty^2(\tau) := \sup_{0\le\gamma\le 1}\sigma_\infty(\tau,\gamma). \tag{5.24}$$

Since we obtain rather complicated expressions for (5.24), we do not give the details.

The next step is to compute the maximum of these two functions, i.e., we define

$$q_{GUESS}(\tau) := \max\{q_0(\tau), q_\infty(\tau)\}$$

and guess that this equals $q_{TG}(\tau)$, defined in (5.23). By construction, $q_{TG}(\tau) \ge q_{GUESS}(\tau)$ holds for all τ, i.e., if $q_{GUESS}(\tau)$ is an upper bound, it is also sharp.

The computation of q_{GUESS} is also done using CAD. Recall that the suprema in (5.24) were computed by solving a quantifier elimination problem. Therefore, $q_0^2(\tau)$ is the smallest λ_0 satisfying a (non-quantified) formula $B_0(\lambda_0,\tau)$ and $q_\infty^2(\tau)$ is the smallest λ_∞ satisfying a (non-quantified) formula $B_\infty(\lambda_\infty,\tau)$. Then $q_{GUESS}^2(\tau)$ is the smallest λ satisfying both formulas, i.e.,

$$B_0(\lambda,\tau) \wedge B_\infty(\lambda,\tau). \tag{5.25}$$

We use Mathematica's command **CylindricalDecomposition** to obtain a representation of the set characterized by (5.25) as a union of cylindrical cells. Using such a representation, the (piecewise polynomial) function $q_{GUESS}^2(\tau)$ can be determined by inspection.

We obtain

$$q^2_{GUESS}(\tau) = \begin{cases} (1-2\tau)^2(2+4(\tau-1)\tau) & \text{for } 0 \leq \tau < \tau_1 \\ \frac{1}{2}(2-\tau) & \text{for } \tau_1 \leq \tau < \tau_2 \\ (1-2\tau)^2(2+4(\tau-1)\tau) & \text{for } \tau_2 \leq \tau \leq 1 \end{cases},$$

where $\tau_1 < \tau_2$ are the two real solutions of

$$(1-2\tau)^2(2+4(\tau-1)\tau) = \frac{1}{2}(2-\tau).$$

To show that $q_{GUESS}(\tau)$ is an upper bound, we set up the quantified formula

$$\forall\, 0 \leq \tau \leq 1 \;:\; \forall\, \eta > 0 \;:\; \forall\, 0 \leq \gamma \leq 1 \;:\; \sigma^2(\tau,\eta,\gamma) \leq q^2_{GUESS}(\tau). \tag{5.26}$$

Since q^2_{GUESS} is a piecewise polynomial function, we may split (5.26) into the intervals $[0,\tau_1)$, $[\tau_1,\tau_2)$ and $[\tau_2,1]$ used in the definition of q^2_{GUESS}. Again we use Mathematica's Resolve command, which reduces these formulas to the logical constant true, i.e, which shows that $q_{GUESS}(\tau)$ is an upper bound. Therefore, we obtain.

Theorem 73 *The convergence rate for the two-grid method using $\nu = \nu_{pre} + \nu_{post} = 1+1$ smoothing steps of collective Jacobi relaxation is given by*

$$q_{TG}(\tau) = \max\left\{|1-2\tau|\sqrt{2+4(\tau-1)\tau}, \frac{1}{4}(\tau-2)^2\right\}$$

for all $\tau \in [0,1]$, which can be seen in Figure 5.4.

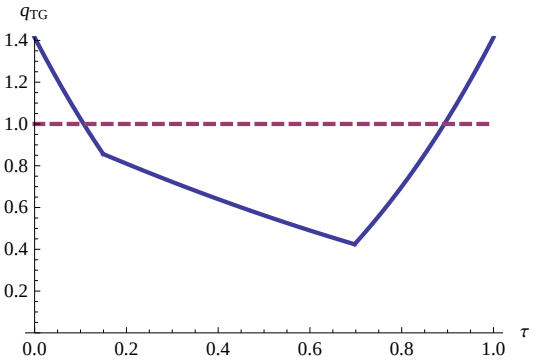


Figure 5.4: Two-grid convergence factor depending on damping parameter

Using this closed form, we can find out for which choices of τ the method converges. Moreover, we obtain that the best choice is $\tau \approx 0.70$ with $q_{TG}(\tau) \approx 0.42$. Still, we should keep in mind, that the computed rate is a sharp worst-case analysis of the convergence rate.

5.5 Summary

In this chapter we have seen that symbolic local Fourier analysis – the combination of local Fourier analysis with cylindrical algebraic decomposition – is a strategy for computing convergence rates or smoothing rates in an entirely automatic manner. The convergence rates of the particular problems and two-grid solvers computed above are not only interesting results on their own but they have the character of model problems for the method of symbolic local Fourier analysis and illustrate the prospects of that method.

Also the analysis for higher dimensions as well as the analysis of other smoothers or the analysis of multigrid methods for other model problems leads to an expression that is a rational function in the mesh size h_k, certain parameters (like α in our case), the damping parameter τ, and trigonometric expressions of the frequencies θ. This is in particular the case for the model problem described in this chapter for the above mentioned generalizations.

Theoretically, all these problems can be solved in finite but not necessarily in reasonable time. Therefore, it is necessary to apply proper strategies to reduce the complexity of the problems as we could see in the last two sections.

Chapter 6

Numerical results

In this chapter we will illustrate the convergence results presented in the last two chapters for model problems. The author will comment on the choice of the damping parameter τ, which can be chosen larger as predicted in Chapter 4.

This chapter is organized as follows. In Section 6.1, we will consider the case $\alpha = 1$. In this case, we will see that for all model problems optimal complexity can be observed. In Section 6.2, we will concentrate on robustness. We will note, as proposed by the convergence theory, that for Model Problem 2, accordingly constructed methods show robust convergence behavior. We will observe such a behavior also in cases where the theory does not state this.

6.1 Optimal complexity

6.1.1 Distributed control model problem

Here, we fix a simple domain $\Omega := (0,1)^2$. The coarsest grid consists of two triangles which are constructed by introducing an edge between the nodes $(0,1)$ and $(1,0)$, cf. Figure 6.1. The refinement is done in an uniform way. First, we give numerical examples for the *distributed control* Model Problem 2.

The model problem reads as follows. Find the control $u \in L^2(\Omega)$ and the state $y \in H^1(\Omega)$ such that they minimize the cost functional J, given by

$$J(y,u) = \frac{1}{2}\|y - y_D\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)},$$

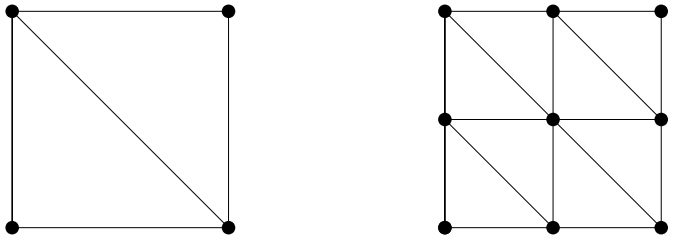


Figure 6.1: Initial mesh and uniform refinement

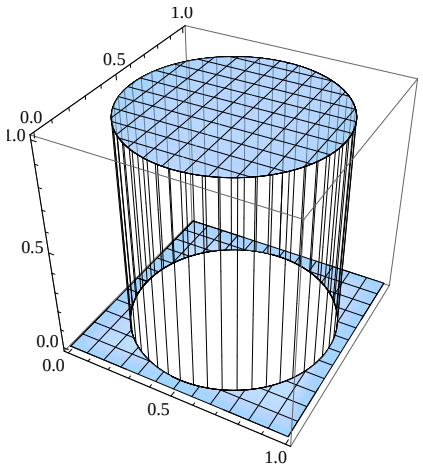


Figure 6.2: Desired state y_D

subject to the elliptic boundary value problem

$$-\Delta y + y = u \text{ in } \Omega \qquad \text{and} \qquad \frac{\partial y}{\partial n} = 0 \text{ on } \partial\Omega,$$

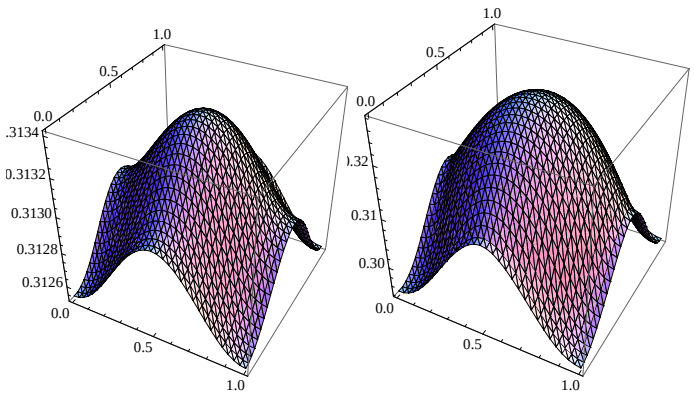
where $y_D : (0,1)^2 \to \mathbb{R}$ is given by

$$y_D(x) := \begin{cases} 1 & \text{if } \left\|x - \left(\frac{1}{2}, \frac{1}{2}\right)\right\|_{\ell^2} \leq \sqrt{\frac{1}{5}} \\ 0 & \text{otherwise} \end{cases},$$

cf. Figure 6.2. As mentioned, here $\alpha := 1$ is fixed. The optimal state and the optimal control are shown in Figure 6.3.

In this thesis we have proposed to use a multigrid method with the preconditioned normal equation smoother (which can be applied to the 3-by-3 formulation of the KKT-system and to the reduced (2-by-2) KKT-system) and the collective point smoothers (which we have only introduced for the reduced (2-by-2) KKT-system).

First we start with the original 3-by-3 formulation of the KKT-system. For this formulation, we can apply the preconditioned normal equation smoother (cf. Subsections 3.2.1



(a) Optimal state (b) Optimal control

Figure 6.3: Distributed control Model Problem 2 for $\alpha = 1$

and 4.1.4) where the matrix $\mathcal{L}_k$ is defined analogously to (4.22). Analogously to the statement in Corollary 33, convergence can be shown for $\tau \in (0, 1/8)$. In fact, this bound is not sharp and the convergence rate can be improved significantly if larger values of τ are chosen, cf. Table 6.1 and Figure 6.4.

Here and in what follows, a W-cycle multigrid method with ν pre- and ν post-smoothing steps was used for simulation. It has to be mentioned that for the model problems, shown here, also the V-cycle converges with rates comparable with the convergence rates of the W-cycle method. The number of iterations and convergence rates were measured as follows: we start with a random initial error and measure the reduction of the error in each step using the norm $\|\cdot\|_{X_{-,k}}$. The iteration was stopped when the initial error was reduced by a factor of $\epsilon = 10^{-6}$. The convergence rates q is the mean convergence rate in this iteration, i.e.,

$$q = \left(\frac{\|x_k^{(n)} - x_k\|_{X_{-,k}}}{\|x_k^{(0)} - x_k\|_{X_{-,k}}} \right)^{1/n},$$

where n is the number of iterations needed to reach the stopping criterion. Here, x_k is the exact solution and $x_k^{(i)}$ is the i-th iterate.

Table 6.1 shows that the best convergence rates are obtained for $\tau = 7/16$. Therefore, we will use this choice for the further calculations. In Table 6.2, we see that the preconditioned normal equation smoother is convergent on all tested grid levels k already for $\nu = 1$ pre- and post-smoothing steps. The convergence theory states that the convergence rates behave like $\nu^{-1/2}$, the numerical results show a faster decay. Moreover, we see that the convergence rates are independent of the grid level k. This means that the overall computational complexity (the number of floating point operations that have to be performed) is proportional to the computational complexity of one multi-

	n	q
$\tau = 1/16$	> 100	0.95
$\tau = 2/16$	> 100	0.92
$\tau = 3/16$	> 100	0.88
$\tau = 4/16$	86	0.85
$\tau = 5/16$	69	0.82
$\tau = 6/16$	57	0.78
$\tau = 7/16$	48	0.75
$\tau = 8/16$	> 100	0.96

Table 6.1: 3-by-3 distributed control: Number of iterations n and convergence rate q for *normal equation smoother* on grid level $k = 5$ and $\nu = \nu_{pre} + \nu_{post} = 1+1$ smoothing steps

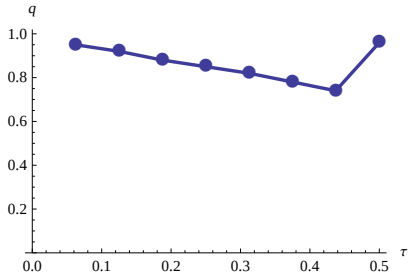


Figure 6.4: 3-by-3 distributed control: convergence rate for *normal equation smoother* on grid level $k = 5$ and $\nu = 4$ smoothing steps

	$\nu = 1+1$		$\nu = 2+2$		$\nu = 4+4$		$\nu = 8+8$	
	n	q	n	q	n	q	n	q
$k = 5$	49	0.75	24	0.56	14	0.35	10	0.24
$k = 6$	49	0.75	25	0.57	14	0.36	10	0.24
$k = 7$	49	0.75	25	0.57	14	0.36	10	0.25
$k = 8$	49	0.75	25	0.57	14	0.36	10	0.25
$k = 9$	49	0.75	25	0.57	14	0.36	10	0.25

Table 6.2: 3-by-3 distributed control: Number of iterations n and convergence rate q for *normal equation smoother* for $\tau = 7/16$

grid cycle. It can be verified that the computational complexity of one multigrid cycle is proportional to the number of unknowns N_k.

An alternative approach is based on the reduction of the problem to a 2-by-2 formulation (reduced KKT-system). For this case we have introduced two kinds of smoothers. We can apply collective point smoothers (cf. Subsection 3.2.2) or we can again use a preconditioned normal equation smoother, which we consider first. Here, the implementation follows (4.14) and again, $\tau = 7/16$ is a good choice for the damping parameter. In Table 6.3, we see that the multigrid method with the preconditioned normal equation smoother is convergent for all tested grid levels k and already for $\nu = \nu_{pre} + \nu_{post} = 1+1$ smoothing steps. Moreover, we observe optimal complexity and again the convergence rate decays faster than $\nu^{-1/2}$ which was proposed by the theory. The convergence rates are comparable with the case of the 3-by-3 formulation.

	$\nu = 1+1$		$\nu = 2+2$		$\nu = 4+4$		$\nu = 8+8$	
	n	q	n	q	n	q	n	q
$k = 5$	47	0.74	24	0.56	14	0.35	10	0.24
$k = 6$	48	0.75	25	0.57	14	0.36	10	0.24
$k = 7$	48	0.75	25	0.57	14	0.36	10	0.25
$k = 8$	49	0.75	25	0.57	14	0.36	10	0.25
$k = 9$	49	0.75	25	0.57	14	0.36	10	0.25

Table 6.3: 2-by-2 distributed control: Number of iterations n and convergence rate q for *normal equation smoother* for $\tau = 7/16$

An alternative approach are collective point smoothers. Here, we stick to the case of the collective Jacobi smoother. Convergence theory shows convergence for $\tau \in (0, 1)$. As due to Table 6.4 (cf. Figure 6.5), $\tau = 3/4$ seems to be optimal, we use that choice. In Figure 6.6, we compare the observed convergence rates with the convergence rates computed in Section 5.4 using local Fourier analysis. There we have analyzed the two-grid method in the one dimensional case. Here, we have used a W-cycle multigrid method for a two-dimensional problem. Nonetheless, we see that the bounds for the convergence rates computed with local Fourier analysis seem to be quite precise estimates for the true behavior.

In Table 6.5 we see that the multigrid method with the collective Jacobi smoother shows an optimal convergence behavior. Moreover, we see that the collective Jacobi smoother leads to much better convergence rates than the preconditioned normal equation smoother. Moreover, one step of the preconditioned normal equation smoother requires approximately twice as much floating point operations than one step of the

	n	q
$\tau = 1/8$	91	0.86
$\tau = 2/8$	44	0.73
$\tau = 3/8$	29	0.61
$\tau = 4/8$	21	0.51
$\tau = 5/8$	16	0.41
$\tau = 6/8$	13	0.33
$\tau = 7/8$	21	0.51
$\tau = 15/16$	44	0.73
$\tau = 8/8$	> 100	0.97

Table 6.4: 2-by-2 distributed control: Number of iterations n and convergence rate q for *collective Jacobi smoother* on grid level $k = 5$ and $\nu = \nu_{pre} + \nu_{post} = 1+1$ smoothing steps

	$\nu = 1+1$		$\nu = 2+2$		$\nu = 4+4$		$\nu = 8+8$	
	n	q	n	q	n	q	n	q
$k = 5$	13	0.34	8	0.16	6	0.08	4	0.03
$k = 6$	13	0.34	8	0.17	6	0.08	5	0.04
$k = 7$	13	0.34	8	0.17	6	0.08	5	0.04
$k = 8$	13	0.34	8	0.17	6	0.08	5	0.04
$k = 9$	13	0.34	8	0.17	6	0.08	5	0.04

Table 6.5: 2-by-2 distributed control: Number of iterations n and convergence rate q for *collective Jacobi smoother* for $\tau = 3/4$

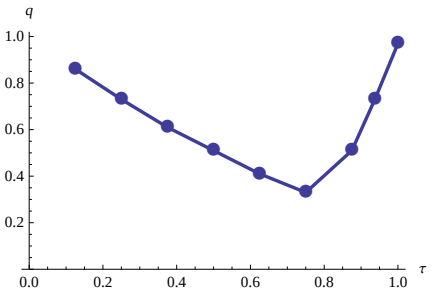


Figure 6.5: 2-by-2 distributed control: convergence rate for *collective Jacobi smoother* on grid level $k = 5$ and $\nu_{pre} + \nu_{post} = 1 + 1$ smoothing steps

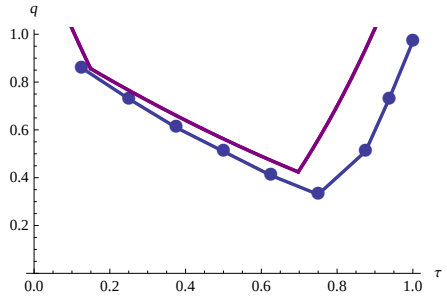


Figure 6.6: 2-by-2 distributed control: convergence rate for *collective Jacobi smoother* compared with result from local Fourier analysis

collective Jacobi smoother, which makes the multigrid method with the collective Jacobi smoother much faster, cf. Table 6.6 for the CPU times on a standard PC. Moreover, we obtain that the CPU time grows linearly with the number of unknowns. Here, for the 3-by-3 formulation the number of unknowns is 3 times the number of nodes and for the 2-by-2 formulation the number of unknowns is 2 times the number of nodes. The methods were realized (including the choice of τ) as outlined above.

	N_k	3-by-3 Normal equation	2-by-2 Normal equation	Collective Jacobi
$k=5$	1 089	0.10 sec	0.05 sec	0.01 sec
$k=6$	4 225	0.35 sec	0.19 sec	0.04 sec
$k=7$	16 641	1.50 sec	0.84 sec	0.18 sec
$k=8$	66 049	6.34 sec	4.01 sec	0.94 sec

Table 6.6: Distributed control: Number of nodes N_k and CPU times for all three approaches for $\nu = \nu_{pre} + \nu_{post} = 1 + 1$ smoothing steps

6.1.2 Boundary control model problem

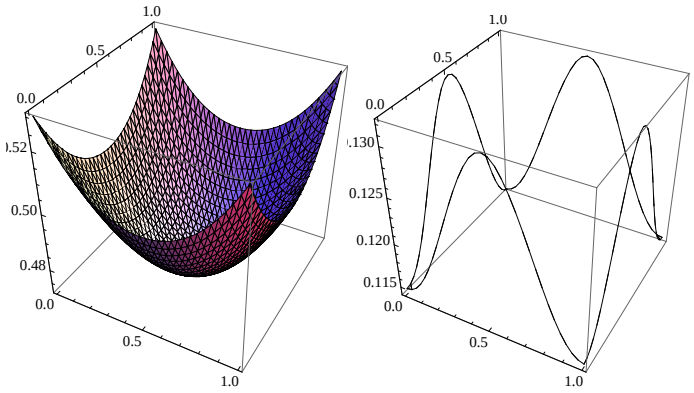
In this subsection, we consider a more advanced problem: the boundary control Model Problem 3. We discuss, how the solution of the model problem looks like, first. The model problem reads as follows. Find control $u \in L^2(\partial\Omega)$ and state $y \in H^1(\Omega)$ such that they minimize the cost functional J, given by

$$J(y,u) = \frac{1}{2}\|y - y_D\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\partial\Omega)},$$

subject to the elliptic boundary value problem

$$-\Delta y + y = 0 \text{ in } \Omega \qquad \text{and} \qquad \frac{\partial y}{\partial n} = u \text{ on } \partial\Omega,$$

where y_D is the same function as in the distributed control case. For $\alpha = 1$, the optimal state and the optimal control are shown in Figure 6.7.



(a) Optimal state (b) Optimal control

Figure 6.7: Boundary control Model Problem 3 for $\alpha = 1$

	3-by-3 Normal equation				2-by-2 Normal equation				Collective Jacobi			
	$\nu = 1+1$		$\nu = 2+2$		$\nu = 1+1$		$\nu = 2+2$		$\nu = 1+1$		$\nu = 2+2$	
	n	q	n	q	n	q	n	q	n	q	n	q
$k = 5$	48	0.75	25	0.57	47	0.74	24	0.56	13	0.34	8	0.16
$k = 6$	48	0.75	24	0.56	48	0.75	25	0.57	13	0.34	8	0.17
$k = 7$	48	0.75	25	0.57	49	0.75	25	0.57	13	0.34	8	0.17
$k = 8$	49	0.75	25	0.57	49	0.75	25	0.57	13	0.34	8	0.17
$k = 9$	49	0.75	25	0.57	49	0.75	25	0.57	13	0.34	8	0.17

Table 6.7: Boundary control: Number of iterations n and convergence rate q for all three proposed methods

Again, we apply the same numerical tests as in the case of the distributed control model problem. Here, we want to mention that the boundary control problem is not covered by the convergence theory for the multigrid method with the collective point smoothers, but it is covered by the convergence theory for the multigrid method with the preconditioned normal equation smoother. We again follow the definition of the matrix, given in (4.14) and (4.22).

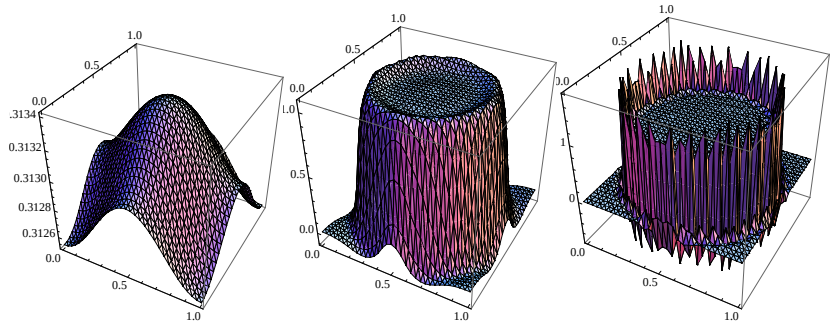
Numerical experiments indicate that the optimal choices of the damping parameter ($\tau = 7/16$ for the preconditioned normal equation smoother and $\tau = 3/4$ for the collective Jacobi smoother) are also optimal choices for the boundary control model problem. The numerical tests are collected in Table 6.7. Again all three approaches shows optimal complexity. Again, the convergence rates decay faster than $\nu^{-1/2}$ and the collective Jacobi smoother is the fastest smoother.

6.2 Robustness

6.2.1 Distributed control model problem

The next step is to analyze the convergence behavior for α approaching 0. Here, we restrict ourselves to the 2-by-2 formulation of the model problem as we have restricted ourselves to that case also for the analysis. Again, we consider the distributed control Model Problem 2 first.

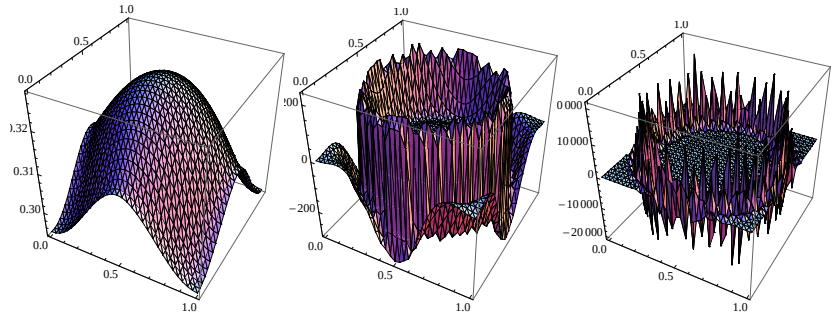
We have computed the solution for various choices of α. For $\alpha = 1$, 10^{-6} and 10^{-12}, the optimal state is shown in Figure 6.8, the optimal control is shown in Figure 6.9. We see that for $\alpha = 10^{-6}$ the optimal state can be seen as a reconstruction of the desired state. This is non-trivial as we have chosen a general L^2-function as desired state which cannot be reconstructed as H^1-function in a trivial way. For the case $\alpha = 10^{-12}$ we observe oscillations at the position where the desired state jumps.



(a) $\alpha = 1$ (b) $\alpha = 10^{-6}$ (c) $\alpha = 10^{-12}$

Figure 6.8: Distributed control Model Problem 2: Optimal state

The methods we have proposed to obtain robust convergence rates, was the multigrid method with the preconditioned normal equation smoother and the collective point



(a) $\alpha = 1$ (b) $\alpha = 10^{-6}$ (c) $\alpha = 10^{-12}$

Figure 6.9: Distributed control Model Problem 2: Optimal control

	$\alpha = 1$		$\alpha = 10^{-4}$		$\alpha = 10^{-8}$		$\alpha = 10^{-12}$	
	n	q	n	q	n	q	n	q
$k = 5$	48	0.75	50	0.76	40	0.71	50	0.76
$k = 6$	48	0.75	48	0.75	51	0.76	53	0.77
$k = 7$	49	0.75	48	0.75	54	0.77	56	0.78
$k = 8$	49	0.75	49	0.75	50	0.76	44	0.73
$k = 9$	49	0.75	49	0.75	49	0.75	44	0.73

Table 6.8: 2-by-2 distributed control: Number of iterations n and convergence rate q for *preconditioned normal equation smoother* for $\tau = 7/16$ and $\nu = \nu_{pre} + \nu_{post} = 1 + 1$ smoothing steps

	$\alpha = 1$		$\alpha = 10^{-4}$		$\alpha = 10^{-8}$		$\alpha = 10^{-12}$	
	n	q	n	q	n	q	n	q
$k = 5$	13	0.34	13	0.33	9	0.21	13	0.33
$k = 6$	13	0.34	13	0.34	12	0.29	13	0.33
$k = 7$	13	0.34	13	0.34	13	0.33	13	0.33
$k = 8$	13	0.34	13	0.34	13	0.34	11	0.26
$k = 9$	13	0.34	13	0.34	13	0.34	10	0.25

Table 6.9: 2-by-2 distributed control: Number of iterations n and convergence rate q for *collective Jacobi smoother* for $\tau = 3/4$ and $\nu = \nu_{pre} + \nu_{post} = 1 + 1$ smoothing steps

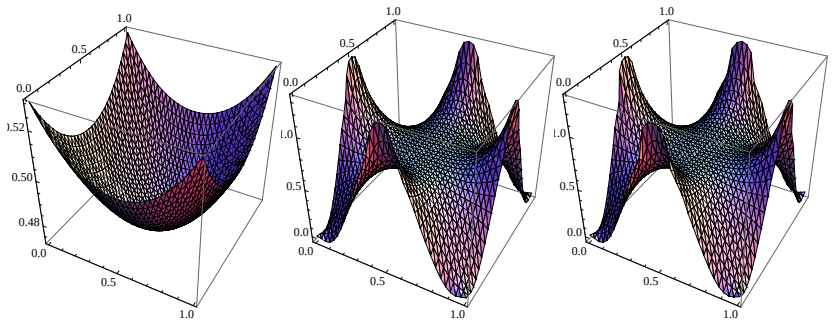
	$\alpha = 1$		$\alpha = 10^{-4}$		$\alpha = 10^{-8}$		$\alpha = 10^{-12}$	
	n	q	n	q	n	q	n	q
$k = 5$	7	0.10	6	0.10	9	0.18	6	0.07
$k = 6$	7	0.10	6	0.10	8	0.17	6	0.07
$k = 7$	7	0.10	7	0.10	7	0.12	6	0.07
$k = 8$	7	0.10	7	0.10	7	0.11	7	0.13
$k = 9$	7	0.10	7	0.10	7	0.10	8	0.17

Table 6.10: 2-by-2 distributed control: Number of iterations n and convergence rate q for *collective Gauss Seidel smoother* for $\nu = \nu_{pre} + \nu_{post} = 1 + 1$ smoothing steps

smoothers. First we give the convergence tables, where the preconditioned normal equation smoother was used, in Table 6.8. We observe that convergence behavior is both, independent of the grid level k and robust in the choice of α. The same convergence behavior can also be observed for the collective Jacobi smoother, see Table 6.9. As already mentioned in Section 3.2.2, we can improve the convergence rates further by using collective Gauss Seidel iteration, cf. Table 6.10. Here, damping is not necessary and therefore we have used an undamped version. We have to mention that convergence theory is not available for collective Gauss Seidel iteration but that method is a canonical extension of collective Jacobi iteration.

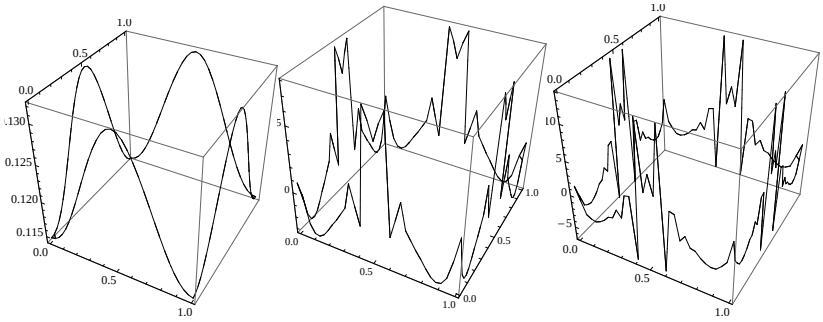
6.2.2 Boundary control model problem

First we want to show how the solution of the model problem looks like. We have computed the solution for various values of α. For $\alpha = 1$, 10^{-6} and 10^{-12}, the optimal state is shown in Figure 6.10, the optimal control is shown in Figure 6.11.



(a) $\alpha = 1$ (b) $\alpha = 10^{-6}$ (c) $\alpha = 10^{-12}$

Figure 6.10: Boundary control Model Problem 3: Optimal state



(a) $\alpha = 1$ (b) $\alpha = 10^{-6}$ (c) $\alpha = 10^{-12}$

Figure 6.11: Boundary control Model Problem 3: Optimal control

	$\alpha = 1$		$\alpha = 10^{-4}$		$\alpha = 10^{-8}$		$\alpha = 10^{-12}$	
	n	q	n	q	n	q	n	q
$k = 5$	13	0.34	13	0.34	16	0.40	16	0.41
$k = 6$	13	0.34	13	0.34	17	0.43	16	0.42
$k = 7$	13	0.34	13	0.34	19	0.48	17	0.44
$k = 8$	13	0.34	13	0.34	18	0.45	17	0.43
$k = 9$	13	0.34	13	0.34	14	0.35	17	0.44

Table 6.11: 2-by-2 boundary control: Number of iterations n and convergence rate q for *collective Jacobi smoother* for $\tau = 3/4$ and $\nu = \nu_{pre} + \nu_{post} = 1 + 1$ smoothing steps

	$\alpha = 1$		$\alpha = 10^{-4}$		$\alpha = 10^{-8}$		$\alpha = 10^{-12}$	
	n	q	n	q	n	q	n	q
$k = 5$	6	0.10	7	0.10	10	0.22	11	0.24
$k = 6$	7	0.10	7	0.10	12	0.27	11	0.27
$k = 7$	7	0.10	7	0.11	12	0.30	11	0.26
$k = 8$	7	0.10	7	0.10	12	0.29	12	0.28
$k = 9$	7	0.10	7	0.10	8	0.14	12	0.29

Table 6.12: 2-by-2 boundary control: Number of iterations n and convergence rate q for *collective Gauss Seidel smoother* for $\nu = \nu_{pre} + \nu_{post} = 1 + 1$ smoothing steps

As mentioned, we do not know a norm $\|\cdot\|_X$ such that the problem is well posed (in the sense of condition **(A1)**, introduced on page 17) robust with respect to α, i.e., such that the constants in **(A1)** are independent of α. Therefore, we cannot construct a norm $\|\cdot\|_{X_{-,k}}$, as outlined in Chapter 4, and therefore we cannot even apply the preconditioned normal equation smoother.

For the collective Jacobi smoother and the collective Gauss Seidel smoother, we do not need to know these norms for applying the method. The numerical experiments show good convergence behavior also for the boundary control case, which is not covered by the convergence theory, cf. Table 6.11 for the collective Jacobi iteration and Table 6.11 for the collective Gauss Seidel iteration.

6.2.3 Distributed control model problem on a non-convex domain

In Section 4.5 we have shown that the methods introduced in this thesis for the distributed control Model Problem 2 also converge on domains where the full H^2-ellipticity cannot be guaranteed. One example, for such a domain is the L-shaped domain $\Omega := (0,2)^2 \backslash [1,2)^2$, cf. Figure 4.1. There, we have only partial elliptic regularity, i.e., we can only show that – provided to have sufficiently smooth data – the solution is a function in $H^{2-s}(\Omega)$, where $s \in [0,1)$ is the regularity parameter of assumption **(R')**, introduced on page 91. For the L-shaped domain, we have used the initial mesh shown in Figure 6.12 and uniform refinement. For the L-shaped domain, we have $s > \frac{1}{3}$.

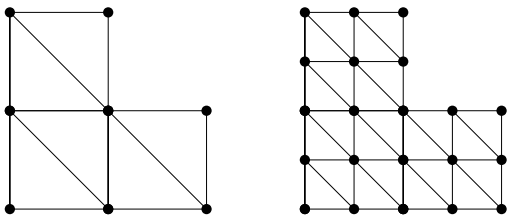


Figure 6.12: Initial mesh and uniform refinement

In the Tables 6.13 and 6.14, we see that the convergence rates for the L-shaped domain and for the domain $\Omega = (0,1)^2$ are comparable, i.e., the multigrid method does not suffer from the lack of full regularity. Convergence theory shows that for varying ν, we cannot expect anymore that the convergence rates behaves like $\nu^{-1/2}$ but like $\nu^{-(1-s)/2}$. However, the numerical experiments show convergence rates that decays even faster than $\nu^{-1/2}$, cf. Table 6.15.

	$\alpha = 1$		$\alpha = 10^{-4}$		$\alpha = 10^{-8}$		$\alpha = 10^{-12}$	
	n	q	n	q	n	q	n	q
$k = 5$	49	0.75	50	0.76	41	0.71	50	0.76
$k = 6$	48	0.75	49	0.75	51	0.76	53	0.77
$k = 7$	49	0.75	49	0.75	55	0.77	56	0.78
$k = 8$	49	0.75	49	0.75	51	0.76	44	0.73
$k = 9$	49	0.75	49	0.75	49	0.75	44	0.73

Table 6.13: 2-by-2 distributed control on L-shaped domain: Number of iterations n and convergence rate q for *preconditioned normal equation smoother* for $\tau = 7/16$ and $\nu = \nu_{pre} + \nu_{post} = 1 + 1$ smoothing steps

	$\alpha = 1$		$\alpha = 10^{-4}$		$\alpha = 10^{-8}$		$\alpha = 10^{-12}$	
	n	q	n	q	n	q	n	q
$k = 5$	14	0.35	13	0.34	9	0.21	13	0.33
$k = 6$	13	0.34	13	0.34	12	0.29	13	0.33
$k = 7$	13	0.34	13	0.34	13	0.33	13	0.33
$k = 8$	13	0.34	13	0.34	13	0.34	11	0.26
$k = 9$	13	0.34	13	0.34	13	0.34	11	0.26

Table 6.14: 2-by-2 distributed control on L-shaped domain: Number of iterations n and convergence rate q for *collective Jacobi smoother* for $\tau = 3/4$ and $\nu = \nu_{pre} + \nu_{post} = 1 + 1$ smoothing steps

	$\nu = 1 + 1$		$\nu = 2 + 2$		$\nu = 4 + 4$	
	n	q	n	q	n	q
Preconditioned normal equation	49	0.75	25	0.57	14	0.36
Collective Jacobi smoother	14	0.35	8	0.17	6	0.08

Table 6.15: 2-by-2 distributed control on L-shaped domain: Number of iterations n and convergence rate q on grid level $k = 5$

Chapter 7

Conclusions

In this thesis we have seen how fast iterative solvers for solving optimality systems arising from optimal control problems can be constructed. Similar to former results, cf. Simon [57] and Schöberl, Simon and Zulehner [53], we could give convergence results based on Hackbusch's splitting into smoothing and approximation property.

In Simon [57], the analysis was restricted to a distributed control model problem (cf. Model Problem 2). The first result we saw in this work was that his convergence results (for α fixed) can be carried over to other model problems. Moreover, in Simon [57], Uzawa type smoothers have been proposed and analyzed. Here, we propose two other kinds of smoothers: on the one hand we propose smoothers based on the normal equation which can be constructed in a more flexible way. Moreover, the analysis of such smoothers is easier than the analysis of Uzawa type smoothers. Smoothers based on the normal equation were known before, but in Takacs and Zulehner [61] we gave numerical results which indicated that they show convergence rates comparable to the convergence rates obtained using Uzawa type smoothers, which was a surprise for us.

On the other hand, we have seen collective point smoothers. Also this kind of smoothers was known, cf. Borzi, Kunisch and Kwak [12]. Here, we have presented a smoothing analysis (cf. Takacs and Zulehner [62]) that guarantees robustness of the convergence rates in α. At a first observation, this result fits into the framework of Schöberl, Simon and Zulehner [53], so we can combine the smoothing analysis with their work to show convergence.

Numerical experiments have shown that the collective point smoothers are much faster than the preconditioned normal equation smoothers. Moreover, even if the convergence analysis is restricted to the distributed control Model Problem 2, collective point smoothers can be applied also to the other model problems, e.g., to the boundary control Model Problem 3. Also in this case, we have observed robust convergence behavior.

Such a result could not be achieved using the normal equation smoothers as insight into the problem is required even for the setup of the method. For applying collective point smoothers in our context no insight is needed.

Under the assumption of full elliptic regularity we gave a convergence proof which is slightly different to the proof given in Schöberl, Simon and Zulehner [53]. This proof can be generalized to domains where full elliptic regularity cannot be guaranteed. Typically, full elliptic regularity requires that the domain has a sufficiently smooth boundary or that the domain is convex. The analysis, we have presented, relaxes this condition and is applicable to any reasonable polygonal (or polyhedral) domain.

Beside qualitative convergence results, we also were interested in quantitative convergence analysis: we wanted to compute sharp upper bounds for the convergence rates. Here, we have used local Fourier analysis. Already in Borzi, Kunisch and Kwak [12], local Fourier analysis has been applied to the problems discussed in this thesis. We could show that a tool from symbolic computation – cylindrical algebraic decomposition (CAD) – allows to compute an explicit representation of the convergence rate as a function of a certain parameter like the damping parameter τ in this thesis (cf. Pillwein and Takacs [48, 49]). In Figure 6.6 we have seen that the results computed with local Fourier analysis are very similar to the cases obtained in numerical experiments.

Bibliography

[1] R. Adams and J. Fournier. *Sobolev Spaces*. Academic Press, 2008. 2nd ed.

[2] E. Arian and S. Ta'asan. Multigrid one-shot methods for optimal control problems: Infinite dimensional control. ICASE-Report 94–52, NASA Langley Research Center, Hampton VA, 1994.

[3] K. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Nonlinear Programming*. Stanford University Press, Stanford, CA, 1958.

[4] I. Babuška. Error-bounds for finite element method. *Numerische Mathematik*, 16(4):322 – 333, 1971.

[5] A. Battermann and M. Heinkenschloss. Preconditioners for Karush-Kuhn-Tucker Matrices Arising in the Optimal Control of Distributed Systems. In W. Desch, F. Kappel, and K. Kunisch, editors, *Control and Estimation of Distributed Parameter Systems*, volume 126 of *International Series of Numerical Mathematics*, pages 15 – 32. Birkhäuser Basel, 1998.

[6] A. Battermann and E.W. Sachs. Block preconditioners for KKT systems in PDE-governed optimal control problems. In Hoffmann, Karl-Heinz (ed.) et al., editor, *Fast solution of discretized optimization problems. Workshop held at the Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany, May 8 – 12, 2000*, pages 1 – 18. Basel: Birkhäuser. ISNM, Int. Ser. Numer. Math. 138, 2001.

[7] M. P. Bendsøe and O. Sigmund. *Topology Optimization: Theory, Methods and Applications*. Springer, Berlin, 2003.

[8] M. Benzi, G.H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1 – 137, 2005.

[9] M. Benzi and A.J. Wathen. Some preconditioning techniques for saddle point problems. In *Model order reduction: theory, research aspects and applications*, volume 13 of *Math. Ind.*, pages 195 – 211. Springer, Berlin, 2008.

[10] G. Biros and O. Ghattas. Parallel Lagrange-Newton–Krylov-Schur methods for PDE-constrained optimization II: The Lagrange-Newton solver and its application to optimal control of steady viscous flows. *SIAM J. on Scientific Computing*, 27(2):714 – 739, 2005.

[11] G. Biros and O. Ghattas. Parallel Lagrange-Newton-Krylov-Schur methods for PDE-constrained optimization I: The Krylov-Schur solver. *SIAM J. on Scientific Computing*, 27(2):687 – 713, 2005.

[12] A. Borzi, K. Kunisch, and D.Y. Kwak. Accuracy and convergence properties of the finite difference multigrid solution of an optimal control optimality system. *SIAM J. on Control and Optimization*, 41(5):1477 – 1497, 2003.

[13] A. Borzi and V. Schulz. Multigrid Methods for PDE Optimization. *SIAM Review*, 51:361 – 395, 2009.

[14] D. Braess. *Finite Elemente*. Springer, 1992.

[15] D. Braess and W. Hackbusch. A New Convergence Proof for the Multigrid Method Including the V-Cycle. *SIAM J. on Numerical Analysis*, 20(5):967 – 975, 1983.

[16] D. Braess and R. Sarazin. An efficient smoother for the Stokes problem. *Applied Numerical Mathematics*, 23(1):3 – 19, 1997.

[17] J.H. Bramble. *Multigrid Methods*. Longman Scientific & Technical, Longman House, Essex, England, 1993.

[18] J.H. Bramble and J.E. Pasciak. A Preconditioning Technique for Indefinite Systems Resulting from Mixed Approximations of Elliptic Problems. *Mathematics of Computation*, 50(181):1 – 17, 1988.

[19] A. Brandt. Multi-level adaptive solutions to boundary-value problems. *Math. Comp.*, 31:333 – 390, 1977.

[20] A. Brandt. Rigorous Quantitative Analysis of Multigrid, I: Constant Coefficients Two-Level Cycle with L_2-Norm. *SIAM J. on Numerical Analysis*, 31(6):1695 – 1730, 1994.

[21] S. Brenner and L. Scott. *The Mathematical Theory of Finite Element Methods.* Springer-Verlag, New York, 1994.

[22] S.C. Brenner. Multigrid methods for parameter dependent problems. *RAIRO, Modélisation Math. Anal. Numér*, 30:265 – 297, 1996.

[23] F. Brezzi. On the Existence, Uniqueness and Approximation of Saddle Point Problems Arising from Lagrangian Multipliers. *RAIRO Anal. Numér.*, 8(2):129 – 151, 1974.

[24] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods.* Springer-Verlag, 1991.

[25] C.W. Brown. QEPCAD B – a program for computing with semi-algebraic sets. *Sigsam Bulletin*, 37(4):97 – 108, 2003.

[26] P. Butzer and H. Berens. *Semi-Groups of Operators and Approximation.* Springer-Verlag, Berlin Heidelberg New York, 1967.

[27] G.E. Collins. Quantifier elimination for real closed fields by cylindrical algebraic decomposition. In *Automata theory and formal languages (Second GI Conf., Kaiserslautern, 1975)*, pages 134 – 183. Lecture Notes in Comput. Sci., Vol. 33. Springer, Berlin, 1975.

[28] M. Crouzeix and P.A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations I. *RAIRO Anal. Numér.*, 7(R-3):33 – 76, 1973.

[29] M. Dauge. Elliptic boundary value problems on corner domains. Smoothness and asymptotics of solutions. *Lecture Notes in Mathematics, 1341. Berlin etc.: Springer-Verlag*, 1988.

[30] M. Dauge. Neumann and mixed problems on curvilinear polyhedra. *Integral Equations Oper. Theory*, 15:227 – 261, 1992.

[31] A. Ecker and W. Zulehner. On the Smoothing Property of Multi-Grid Methods in Non-Symmetric Case. *Numerical Linear Algebra with Applications*, 3(2):161 – 172, 1996.

[32] H. Gfrerer. Generalized penalty methods for a class of convex optimization problems with pointwise inequality constraints. NuMa-Report 6, 2009.

[33] Grisvard, P. *Singularities in Boundary Value Problems.* Springer-Verlag, Berlin Heidelberg New York London, 1992.

[34] W. Hackbusch. Fast solution of elliptic control problems. *J. Optimization Theory Appl.*, 31:565 – 581, 1980.

[35] W. Hackbusch. *Multi-Grid Methods and Applications.* Springer, Berlin, 1985.

[36] S.B. Hazra and V. Schulz. Simultaneous pseudo-timestepping for PDE-model based optimization problems. *BIT*, 44(3):457 – 472, 2004.

[37] R. Herzog and E. Sachs. Preconditioned Conjugate Gradient Method for Optimal Control Problems with Control and State Constraints. *SIAM J. on Matrix Anal. & Appl*, 31(5):2291 – 2317, 2010.

[38] M.R. Hestenes and E. Stiefel. Methods of Conjugate Gradients for Solving Linear Systems 1. *Journal Of Research Of The National Bureau Of Standards*, 49(6):409 – 436, 1952.

[39] H. Hong, R. Liska, and S. Steinberg. Applications of quantifier elimination (Albuquerque, NM, 1995). *J. Symbolic Comput.*, 24(2):161 – 187, 1997.

[40] M. Ito and K. Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems and Control Letters*, 50:221 – 228, 2003.

[41] O. Lass, M. Vallejos, A. Borzi, and C.C. Douglas. Implementation and analysis of multigrid schemes with finite elements for elliptic optimal control problems. *Computing*, 84(1 – 2):27 – 48, 2009.

[42] J.L. Lions. *Optimal control of systems governed by partial differential equations.* Berlin-Heidelberg-New York: Springer-Verlag, 1971.

[43] J.L. Lions and Magenes. *Non-Homogeneous Boundary Value Problems and Applications.* Berlin-Heidelberg-New York: Springer-Verlag, 1972.

[44] J. Necas. *Les méthodes directes en théorie des équations elliptiques.* Masson, Paris, 1967.

[45] C. Paige and M.A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. on Numerical Analysis*, 12:617 – 629, 1975.

[46] C. Pearcy. An Elementary Proof of the Power Inequality for the Numerical Radius. *Michigan Math. Journal*, 13(3):289 – 291, 1966.

[47] J.W. Pearson and A.J. Wathen. A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. *Numerical Linear Algebra with Applications*, page n/a, 2011.

[48] V. Pillwein and S. Takacs. Smoothing analysis of an all-at-once multigrid approach for optimal control problems using symbolic computation. In U. Langer and P. Paule, editors, *Numerical and Symbolic Scientific Computing: Progress and Prospects*. Springer, Wien, 2011.

[49] V. Pillwein and S. Takacs. A local Fourier convergence analysis of a multigrid method using symbolic computation, 2012. Submitted (DK-report 04-2012).

[50] O. Pironneau. *Optimal Shape Design for Elliptic Systems*. Springer Series in Computational Physics. Springer, New York, 1984.

[51] T. Rees, S. Dollar, and A. Wathen. Optimal Solvers for PDE-Constrained Optimization. *SIAM J. on Scientific Computing*, 32(1):271 – 298, 2010.

[52] A. Reusken. A new lemma in multigrid convergence theory. *Report-RANA*, 91 – 07, 1991.

[53] J. Schöberl, R. Simon, and W. Zulehner. A Robust Multigrid Method for Elliptic Optimal Control Problems. *SIAM J. on Numerical Analysis*, 49:1482 – 1503, 2011.

[54] J. Schöberl and W. Zulehner. Symmetric Indefinite Preconditioners for Saddle Point Problems with Applications to PDE-Constrained Optimization Problems. *SIAM J. on Matrix Anal. & Appl*, 29(3):752 – 773, 2007.

[55] V. Schulz and G. Wittum. Transforming smoothers for PDE constrained optimization problems. *Computing and Visualization in Science*, 11(4 – 11):207 – 219, 2008.

[56] A. Seidl and T. Sturm. A generic projection operator for partial cylindrical algebraic decomposition. In *Proceedings of the 2003 International Symposium on Symbolic and Algebraic Computation*, pages 240 – 247 (electronic), New York, 2003. ACM.

[57] R. Simon. *Multigrid solvers for saddle point problems in PDE-constrained optimization.* PhD thesis, Johannes Kepler University Linz, SFB013 Numerical and Symbolic Scientific Computing, 2008.

[58] R. Simon and W. Zulehner. On Schwarz-type smoothers for saddle point problems with applications to PDE-constrained optimization problems. *Numerische Mathematik*, 111:445 – 468, 2009.

[59] A. Strzeboński. Solving systems of strict polynomial inequalities. *J. Symbolic Comput.*, 29(3):471 – 480, 2000.

[60] S. Ta'asan. "One-shot" methods for optimal control of distributed parameter systems I: The finite dimensional case. ICASE-Report 91–2, NASA Langley Research Center, Hampton VA, 1991.

[61] S. Takacs and W. Zulehner. Multigrid Methods for Elliptic Optimal Control Problems with Neumann Boundary Control. In *Numerical Mathematics and Advanced Applications 2009: Proceedings of ENUMATH*, pages 855 – 863. Springer, 2010.

[62] S. Takacs and W. Zulehner. Convergence Analysis of Multigrid Methods with Collective Point Smoothers for Optimal Control Problems. *Computing and Visualization in Science*, 14(3):131–141, 2011.

[63] S. Takacs and W. Zulehner. Convergence analysis of all-at-once multigrid methods for elliptic control problems under partial elliptic regularity, 2012. Submitted (DK-report 08-2012).

[64] A. Tarski. *A decision method for elementary algebra and geometry.* University of California Press, Berkeley and Los Angeles, Calif., 1951. 2nd ed.

[65] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen, Theorie, Verfahren und Anwendungen.* Vieweg, Wiesbaden, 2005.

[66] U. Trottenberg, C. Oosterlee, and A. Schüller. *Multigrid.* Academic Press, London, 2001.

[67] S. P. Vanka. Block-implicit multigrid solution of Navier-Stokes equations in primitive variables. *Math. Comp.*, 65:138 – 158, 1986.

[68] R. Wienands and W. Joppich. *Practical Fourier analysis for multigrid methods.* Chapman & Hall/CRC, 2005.

[69] G. Wittum. Multi-grid methods for Stokes and Navier-Stokes equations. Transforming smoothers: Algorithms and numerical results. *Numerische Mathematik*, 54(5):543 – 563, 1988.

[70] G. Wittum. On the convergence of multi-grid methods with transforming smoothers. *Numerische Mathematik*, 57(1):15 – 38, 1990.

[71] W. Zulehner. Non-standard Norms and Robust Estimates for Saddle Point Problems. *SIAM J. on Matrix Anal. & Appl*, 32:536 – 560, 2011.

Curriculum vitae

Stefan Takacs

E-mail:	stefan.takacs@dk-compmath.jku.at
Homepage:	http://www.dk-compmath.jku.at/people/stakacs
Birth:	August 26th, 1984

University and School

Oct 2008 – present	**Research** Assistant (PhD student) Doctoral Program "Computational Mathematics" University of Linz Project title: *Efficient solvers for KKT-systems*
Dec 2006 – Sept 2008	**Master Program** Industrial Mathematics University of Linz Master thesis: *Strategies to optimize a test-program based on the Load Matrix method* Supervisor: H. Gfrerer, Institute for Numerical Mathematics In cooperation with AVL List, Graz
Oct 2003 – Dec 2006	**Bachelor Program** Technical Mathematics University of Linz
Sep 1995 – Jul 2003	**Bundesrealgymnasium** Bundesrealgymnasium Ramsauerstraße, Linz
Sep 1990 – Jul 1995	**Volksschule** Including one year of "Vorschule" Volksschule 50, Linz

Conferences, Talks, Workshops and Research Stays (selected)

May – Aug 2012	**Research Stay** Group of R. Herzog University of Chemnitz, Germany.
Mar 2012	**Invited to Minisymposium** GAMM Annual Scientific Conference Title: *A multigrid framework applied to an elliptic optimal control problem with reduced regularity* Darmstadt, Germany
Sep 2011	**Contributed Talk** ENUMATH conference Title: *Symbolic Local Fourier Analysis for Multigrid Methods with Applications in Optimal Control* Leicester, United Kingdom
Aug – Sep 2011	**Research Stay** Group of V. Schulz University of Trier, Germany.
May 23rd, 2011	**Hearing** Austrian Science Fund (FWF) Evaluation of the doctoral program Vienna, Austria.
Apr 2011	**Contributed Talk** GAMM Annual Scientific Conference Title: *Using symbolic methods to analyze convergence properties of multigrid methods* Graz, Austria
Oct – Dec 2010	**Research Stay** Group of A. Wathen University of Oxford, United Kingdom

Sep 2010	**Contributed Talk** European Multigrid Conference (EMG) Title: *Point Smoothers for Elliptic Optimal Control Problems* Ischia, Italy
Jun, Jul 2009	**Contributed Talk** ENUMATH conference Uppsala, Sweden
Jun 2009	**Summer School** Course on "Isogeometric analysis" DTU Copenhagen Lyngby, Denmark
Summer 2007	**Participation** ECMI Modeling Week INSA Rouen, France

Work and Teaching

Feb 2010	**Project week** "Projektwoche Angewandte Mathematik" For high school students Zell an der Pram, Austria
Oct 2008 – Jul 2010, Oct 2011/12 – Jul 2012	**Lecturer** Tutorial for a basic course in mathematics ("Übung") Fachhochschule Wels, Austria
Winter terms 2006/07, 2007/08	**Student Tutor** Tutorium Numerik für MechatronikerInnen University of Linz, Austria
Summer 2005 – present	**Software development** Development of a statistics tool to evaluate demonstrable reliability of technical systems using the Load Matrix method University of Linz, Austria

Feb 2004 – Jul 2008	**Civilian service and part-time work** Evangelisches Diakoniewerk, Gallneukirchen, Austria
Summers 2002 and 2003	**Work in holidays** Data maintenance and ad-hoc-optimization of production lines BMW Motoren, Steyr, Austria

List of Publications

[1] S. Takacs and W. Zulehner. Multigrid Methods for Elliptic Optimal Control Problems with Neumann Boundary Control. In *Numerical Mathematics and Advanced Applications 2009: Proceedings of ENUMATH*, pages 855 – 863. Springer, 2010.

[2] V. Pillwein and S. Takacs. Smoothing analysis of an all-at-once multigrid approach for optimal control problems using symbolic computation. In U. Langer and P. Paule, editors, *Numerical and Symbolic Scientific Computing: Progress and Prospects*. Springer, Wien, 2011.

[3] S. Takacs and W. Zulehner. Convergence Analysis of Multigrid Methods with Collective Point Smoothers for Optimal Control Problems. *Computing and Visualization in Science*, 14(3):131–141, 2011.

[4] V. Pillwein and S. Takacs. A local Fourier convergence analysis of a multigrid method using symbolic computation. 2012. Submitted (DK-report 04-2012).

[5] S. Takacs and W. Zulehner. Convergence analysis of all-at-once multigrid methods for elliptic control problems under partial elliptic regularity. 2012. Submitted (DK-report 08-2012).